(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 August 2002 (01.08.2002)

PCT

(10) International Publication Number
WO 02/059322 A2

(51) International Patent Classification[7]: C12N 15/52, C12P 21/04

(21) International Application Number: PCT/US01/32354

(22) International Filing Date: 17 October 2001 (17.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/240,879	17 October 2000 (17.10.2000)	US
60/272,207	28 February 2001 (28.02.2001)	US
60/310,385	6 August 2001 (06.08.2001)	US

(71) Applicants and
(72) Inventors: **MIAO, Vivian, Pak, Woon** [CA/CA]; 13750 31 Avenue, Surrey, British Columbia V4P 2B7 (CA). **BRIAN, Paul** [GB/CA]; 697 Moberly Road, Vancouver, British Columbia V5Z 4A4 (CA). **BALTZ, Richard, H.** [US/US]; 6438 North Olney Road, Indianapolis, IN 46220 (US). **SILVA, Christopher, J.** [US/CA]; 1306 Bidwell, Suite 33, Vancouver, British Columbia V6G 2L1 (CA).

(74) Agents: **HALEY, James, F., Jr.** et al.; Fish & Neave, 1251 Avenue of the Americas, New York, NY 10020 (US).

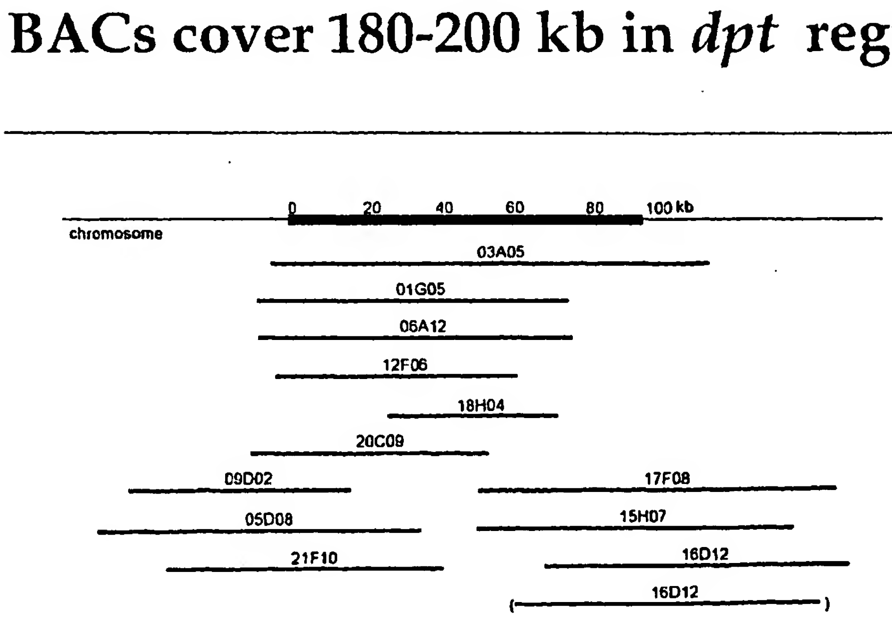
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: COMPOSITIONS AND METHODS RELATING TO THE DAPTOMYCIN BIOSYNTHETIC GENE CLUSTER



(57) Abstract: The invention provides nucleic acid molecules comprising all or a part of a daptomycin biosynthetic gene cluster. The daptomycin biosynthethic gene cluster may be derived from *Streptomyces*, preferably from *S. roseosporus*. The invention also provides other nucleic acid molecules from *S. roseosporus*. The invention further provides polypeptides encoded by the nucleic acid molecules, antibodies that specifically bind to the polypeptides, and methods of using the nucleic acid molecules, polypeptides and antibodies to produce daptomycin and other compounds.

COMPOSITIONS AND METHODS RELATING TO THE DAPTOMYCIN BIOSYNTHETIC GENE CLUSTER

BACKGROUND OF THE INVENTION

Bacteria, including actinomycetes, and fungi synthesize a diverse array of low molecular weight peptide and polyketide compounds (approx. 2-48 residues in length). The biosynthesis of these compounds is catalyzed by non-ribosomal peptide synthetases (NRPSs) and by polyketide syntheses (PKSs). The NRPS process, which does not involve ribosome-mediated RNA translation according to the genetic code, is capable of producing peptides that exhibit enormous structural diversity, compared to peptides translated from RNA templates by ribosomes. These include the incorporation of D- and L-amino acids and hydroxy acids; variations within the peptide backbone which form linear, cyclic or branched cyclic structures; and additional structural modifications, including oxidation, acylation, glycosylation, N-methylation and heterocyclic ring formation. Many non-ribosomally synthesized peptides have been found which have useful pharmacological (e.g., antibiotic, antiviral, antifungal, antiparasitic, siderophore, cytostatic, immunosuppressive, anti-cholesterolemic and anticancer), agrochemical or physicochemical (e.g., biosurfactant) properties.

Non-ribosomally synthesized peptides are assembled by large (e.g., about 200-2000 kDa), multifunctional NRPS enzyme complexes comprising one or more subunits. Examples include daptomycin, vancomycin, echinocandin and cyclosporin. Likewise, polyketides are assembled by large multifunctional PKS enzyme complexes comprising one or more subunits. Examples include erythromycin, tylosin, monensin and avermectin. In some cases, complex molecules can be synthesized by mixed PKS/NRPS systems. Examples include rapamycin, bleomycin and epothilone.

An NRPS usually consists of one or more open reading frames that make up an NRPS complex. The NRPS complex acts as a protein template, comprising a series of protein biosynthetic units configured to bind and activate specific building block substrates and to catalyze peptide chain formation and elongation. (See, e.g., Konz and Marahiel, Chem. Biol., 6, pp. 39-48 (1999) and references cited therein; von Döhren et al., Chem. Biol., 6, pp. 273-279, (1999) and references cited therein; and Cane and Walsh, Chem. Biol., 6, pp. 319-325, (1999), and references cited therein – each hereby incorporated by reference in its entirety). Each NRPS or NRPS subunit comprises one or modules. A "module" is defined as the catalytic unit that incorporates a single building block (e.g., an amino acid) into the growing peptide chain. The order and specificity of the biosynthetic modules that form the NRPS protein template dictates the sequence and structure of the ultimate peptide products.

Each module of an NRPS acts as a semi-autonomous active site containing discrete, folded protein domains responsible for catalyzing specific reactions required for peptide chain elongation. A minimal module (in a single module complex) consists of at least two core domains: 1) an adenylation domain responsible for activating an amino acid (or, occasionally, a hydroxy acid); and 2) a thiolation or acyl carrier domain responsible for transferring activated intermediates to an enzyme-bound pantetheine cofactor. Most modules also contain 3) a condensation domain responsible for catalyzing peptide bond formation between activated intermediates. See Figure 9. Supplementing these three core domains are a variable number of additional domains which can mediate, e.g., N-methylation (M or methylation domain) and L- to D-conversion (E or epimerization domain) of a bound amino acid intermediate, and heterocyclic ring formation (Cy or cyclization domain). The domains are usually characterized by specific amino acid motifs or features. It is the combination of such auxiliary domains acting locally on tethered intermediates within nearby modules that contributes to the enormous structural and functional diversity of the mature peptide products assembled by NRPS and mixed NRPS/PKS enzyme complexes.

The adenylation domain of each minimal module catalyzes the specific recognition and activation of a cognate amino acid. In this early step of non-ribosomal peptide biosynthesis, the cognate amino acid of each NRPS module is bound to the adenylation domain and activated as an unstable acyl adenylate (with concomitant ATP-hydrolysis). See, e.g., Stachelhaus et al., Chem. Biol. 6, pp. 493-505 (1999) and Challis et al., Chem. Biol. 7, pp. 211-224 (2000), each incorporated herein by reference in its entirety. In most NRPS modules, the acyl adenylate intermediate is next transferred to the T (thiolation) domain (also referred to as a peptidyl carrier protein or PCP domain) of the module where it is converted to a thioester intermediate and tethered via a transthiolation reaction to a covalently bound enzyme cofactor (4'-phosphopantetheinyl (4'-PP) intermediate). Modules responsible for incorporating D-configured or N-methylated amino acids may have extra editing domains which, in several NRPSs studied, are located between the A and T domains.

The enzyme-bound thioesterified intermediates in each module are then assembled into the peptide product by stepwise condensation reactions involving transfer of the thioester-activated carboxyl group of one residue in one module to, e.g., the adjacent amino group of the next amino acid in the next module while the intermediates remain linked covalently to the NRPS. Each condensation reaction which mediates peptide chain elongation is catalyzed by a condensation (C) domain which is usually positioned between two modules. The number of condensation domains in a NRPS generally corresponds to the number of peptide bonds present in the final (linear) peptide. An extra C domain has been found in several NRPSs (e.g., at the amino terminus of cyclosporin synthetase and the carboxyl terminus of rapamycin; see, e.g., Konz and Marahiel, *supra*) which has been proposed to be involved in peptide chain termination and cyclization reactions. Many other NRPS complexes, however, release the full length chain in a reaction catalyzed by a C-terminal thioesterase (Te) domain (of approximately 28K-35K relative molecular weight).

Thioesterase domains of most NRPS complexes use a catalytic triad (similar to that of the well-known chymotrypsin mechanism) which includes a conserved serine (less often a cysteine or aspartate) residue in a conserved three-dimensional configuration relative to a histidine and an acidic residue. See, e.g. V. De Crecy-

Lagard in *Comprehensive Natural Products Chemistry*, Volume 4, ed. J.W. Kelly (New York: Elsevier), 1999, pp. 221-238, each incorporated herein by reference in its entirety. Thioester cleavage is a two step process. In the first (acylation) step, the full length peptide chain is transferred from the thiol tethered enzyme intermediate in the thiolation domain (see above) to the conserved serine residue in the Te domain, forming an acyl-O-Te ester intermediate. In the second (deacylation) step, the Te domain serine ester intermediate is either hydrolyzed (thereby releasing a linear, full length product) or undergoes cyclization, depending on whether the ester intermediate is attacked by water (hydrolysis) or by an activated intramolecular nucleophile (cyclization).

Sequence comparisons of C-terminal thioesterase domains from diverse members of the NRPS superfamily have revealed a conserved motif comprising the serine catalytic residue (GXSXG motif), often followed by an aspartic acid residue about 25 amino acids downstream from the conserved serine residue. A second type of thioesterase, a free thioesterase enzyme, is known to participate in the biosynthesis of some peptide and polyketide secondary metabolites. See e.g., Schneider and Marahiel, Arch. Microbiol., 169, pp. 404-410 (1998), and Butler et al., Chem. Biol., 6, pp. 87-292 (1999), each incorporated herein by reference in its entirety. These thioesterases are often required for efficient natural product synthesis. Butler et al. have postulated that the free thioesterase found in the polyketide tylosin gene cluster -- which is required for efficient tylosin production -- may be involved in editing and proofreading functions.

The modular organization of the NRPS multienzyme complex is mirrored at the level of the genomic DNA encoding the modules. The organization and DNA sequences of the genes encoding several different NRPSs have been studied. (See, e.g., Marahiel, Chem. Biol., 4, pp. 561-567 (1997), incorporated herein by reference in its entirety). Conserved sequences characterizing particular NRPS functional domains have been identified by comparing NRPS sequences derived from many diverse organisms and those conserved sequence motifs have been used to design probes useful for identifying and isolating new NRPS genes and modules.

The modular structures of PKS and NRPS enzyme complexes can be exploited to engineer novel enzymes having new specificities by changing the numbers and positions of the modules at the DNA level by genetic engineering and recombination *in vivo*. Functional hybrid NRPSs have been constructed, for example, based on whole-module fusions. See, e.g., Gokhale et al., Science, 284, pp. 482-485 (1999); Mootz et al., Proc. Natl. Acad. Sci. U.S.A., 97, pp. 5848-5853 (2000), incorporated herein by reference in their entirety. Recombinant techniques may be used to successfully swap domains originating from a heterologous PKS or NRPS complex. See, e.g., Schneider et al., Mol. Gen. Genet., 257, pp. 308-318 (1998); McDaniel et al., Proc. Natl. Acad. Sci. U.S.A., 96, pp. 1846-1851 (1999); United States Patent Nos. 5,652,116 and 5,795,738; and International Publication WO 00/56896; incorporated herein by reference in their entirety.

Engineering a new substrate specificity within a module by altering residues which form the substrate binding pocket of the adenylation domain has also been described. See, e.g., Cane and Walsh, Chem. Biol., 6, 319-325 (1999); Stachelhaus et al., Chem. Biol., 6, 493-505 (1999); and WO 00/52152; each incorporated herein by reference in its entirety. By comparing the sequence of the *B. subtilis* peptide synthetase GrsA adenylation domain (PheA) (whose structure is known) with sequences of 160 other adenylation domains from pro- and eukaryotic NPRSs, for example, Stachelhaus et al. (*supra*) and Challis et al., Chem. Biol., 7, pp. 211-224 (2000) defined adenylation (A) domain signature sequences (analogous to codons of the genetic code) for a variety of amino acid substrates. From the collection of those signature sequences, a putative NRPS selectivity-conferring code (with degeneracies like the genetic code) was formulated.

The ability to engineer NRPSs having new modular template structures and new substrate specificities by adding, deleting or exchanging modules (or by adding, deleting or exchanging domains within one or more modules) will enable the production of novel peptides having altered and potentially advantageous properties. A combinatorial library comprising over 50 novel polyketides, for example, was prepared by systematically modifying the PKS that synthesizes an erythromycin precursor (DEBS) by substituting counterpart sequences from the rapamycin PKS (which encodes alternative substrate specificities). See, e.g., WO 00/63361 and McDaniel et al., (1999), *supra*, each incorporated herein by reference in its entirety.

A number of bacteria that produce antibiotics and other potentially toxic compounds synthesize ATP-binding cassette (ABC) transporters. ABC transporters use proton-dependent transmembrane electrochemical potential to export toxic cellular metabolites such as antibiotics, and to import materials from the environment, e.g. iron or other metals. There are three types of ABC transporters and genes encoding pumps responsible for antibiotic resistance, and they are often linked to the biosynthetic cluster in antibiotic producer organisms (e.g. actinorhodin resistance in *Streptomyces coelicolor*). See, e.g., Mendez *et al.*, *FEMS Microbiol. Lett.* 158: 1-8 (1998), herein incorporated by reference. All have ATP-binding regions that include Walker A and B motifs. *Id.* Type I systems involve separate genes for a hydrophilic ATP-binding domain and a hydrophobic integral membrane domain. Type III systems involve a single gene encoding a protein with a hydrophobic N-terminus and a hydrophilic, ATP-binding C-terminus. Type II transporters have no hydrophobic domain, and two sets of Walker motifs, in the order A:B:A:B.

The *Streptomyces glaucescens* genes, StrV (PIR Accession No. S57561) and StrW (PIR Accession No. S57562) encode type III transporters associated with resistance to streptomycin-related compounds. Both genes are within a 5'-hydroxystreptomycin antibiotic biosynthetic gene cluster. See, e.g., Beyer *et al.*, *Mol. Gen. Genet.* 250: 775-84 (1996), herein incorporated by reference. Resistance to doxorubicin and related antibiotics is conferred by two type I transporters in *Streptomyces peucetius*, which are encoded by *drrA* and *drrB*. See, e.g., Guifoile *et al.*, *Proc. Natl. Acad. Sci. USA* 88:8553-57 (1991), herein incorporated by reference. Further, homologs of *drr*AB isolated from *Streptomyces rochei* confer multidrug resistance when expressed under control of the actinorhodin PKS promoter in *S. lividans*. See, e.g., Fernandez-Moreno *et al.*, *J. Bacteriol.* 179: 6929-36 (1998), herein incorporated by reference.

Daptomycin (described by R.H. Baltz in *Biotechnology of Antibiotics*, 2nd Ed., ed. W.R. Strohl (New York: Marcel Dekker, Inc.), 1997, pp. 415-435) is an example of a non-ribosomally synthesized peptide made by a NRPS. Daptomycin, also known as LY146032, is a cyclic lipopeptide antibiotic that is produced by the fermentation of *Streptomyces roseosporus*. Daptomycin is a member of the factor A-21978C type antibiotics of *S. roseosporus* and comprises an n-decanoyl side chain linked via a three-amino acid chain to the N-terminal tryptophan of a cyclic 10-amino acid peptide. The compound is being developed in a variety of formulations to treat serious infections for which therapeutic options are limited, such as infections caused by bacteria including, but not limited to, methicillin resistant *Staphylococcus aureus*, vancomycin resistant enterococci, glycopeptide intermediary susceptible *Staphylococcus aureus*, coagulase-negative staphylococci, and penicillin-resistant *Streptococcus pneumoniae*. See, e.g., Tally *et al., Exp. Opin. Invest. Drugs 8*:1223-1238, 1999. The antibiotic action of daptomycin against Gram-positive bacteria has been attributed to its ability to interfere with membrane potential and to inhibit lipoteichoic acid synthesis.

Identification of the genes encoding the proteins involved in the daptomycin biosynthetic pathway, including the daptomycin NRPS, will provide a first step in producing modified *Streptomyces roseosporus* as well as other host strains which can produce an improved antibiotic (for example, having greater potency); which can produce natural or new antibiotics in increased quantities; or which can produce other peptide products having useful biological properties. Compositions and methods relating to the *Streptomyces roseosporus* daptomycin biosynthetic gene cluster, including isolated nucleic acids and isolated proteins, are described in United States Provisional Applications 60/240,879, filed October 17, 2000; 60/272,207, filed February 28, 2001; and 60/310,385, filed August 8, 2001; all of which are hereby incorporated by reference in its entirety.

It would be advantageous, moreover, to identify the genetic and modular organization of the *Streptomyces roseosporus* daptomycin biosynthetic gene cluster in order to construct full length daptomycin NRPS templates for expression in *Streptomyces roseosporus* and in heterologous hosts. In particular, it would be advantageous to know whether the daptomycin gene cluster comprises a thioesterase (Te) domain. If so, that Te domain could be isolated and used to catalyze peptide chain termination in new NRPS modules and templates by expression as a fusion or as a free peptide. See, e.g., de Ferra et al., J. Biol. Chem., 272, pp. 25304-25309 (1997);

Guenzi et al., J. Biol. Chem., 273, pp. 14403-14410 (1998); and Trauger et al., Nature, 407, pp. 215-218 (2000); each incorporated herein by reference in its entirety. It would also be advantageous to identify other nucleic acid molecules that encode polypeptides involved in daptomycin biosynthesis. These include, without limitation, enzymes involved in attaching a lipid tail to the peptide domain of daptomycin, polypeptides that regulate antibiotic resistance and ABC transporters. Polypeptides that regulate antibiotic resistance and ABC transporters could be used to confer resistance or increase, modify or decrease resistance of a bacteria to daptomycin and related antibiotics. Polypeptides involved in antibiotic resistance would also be useful to determine bacterial mechanisms of resistance, so that daptomycin and related antibiotics can be modified to make them more potent against resistant bacteria.

SUMMARY OF THE INVENTION

The instant invention addresses these problems by providing a nucleic acid molecule that comprises all or a part of a daptomycin biosynthetic gene cluster, preferably one from *S. roseosporus*. The nucleic acid molecule may encode DptA, DptB, DptC or DptD or may comprise one or more of the *dptA*, *dptB*, *dptC* or *dptD* genes from the daptomycin biosynthetic gene cluster of *S. roseosporus*.

The instant invention also provides nucleic acid molecules encoding a free thioesterase and an integral thioesterase from a daptomycin biosynthetic gene cluster. The nucleic acid molecule may encode DptH or the thioesterase domain from DptD, or may comprise the *dptH* or *dptH* gene from the daptomycin biosynthetic gene cluster.

Another object of the invention is to provide a nucleic acid molecule comprising a DNA sequence from a bacterial artificial chromosome comprising a nucleic acid sequence from *S. roseosporus*. The nucleic acid molecule preferably comprises a *S. roseosporus* nucleic acid sequence from any one of bacterial artificial chromosome (BAC) clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05. In a preferred embodiment, the nucleic acid molecule encodes a polypeptide. In another preferred embodiment, the nucleic acid molecule encodes a polypeptide that is involved in daptomycin biosynthesis, such as a *dptA, dptB, dptC, dptD, dptE, dptF, dptH,* an

ABC transporter, or a polypeptide that regulates antibiotic resistance, as described herein.

The invention also provides selectively hybridizing or homologous nucleic acid molecules of the above-described nucleic acid molecules. The invention further provides allelic variants and parts thereof. The invention further provides nucleic acid molecules that comprise one or more expression control sequences controlling the transcription of the above-described nucleic acid molecules. The expression control sequence may be derived from the expression control sequences of the daptomycin biosynthetic gene cluster or may be derived from a heterologous nucleic acid sequence.

In another embodiment, the invention provides a nucleic acid molecule comprising one or more expression control sequences from a gene comprising a nucleic acid sequence that encodes a thioesterase and/or a daptomycin NRPS from the daptomycin biosynthetic gene cluster. Preferably, the nucleic acid molecule comprises a part or all of the expression control sequences of the daptomycin NRPS or *dptH*.

Another object of the invention is to provide a vector and/or host cell comprising one or more of the above-described nucleic acid molecules. In a preferred embodiment, the vector and/or host cell comprises a nucleic acid molecule encoding all or part of DptA, DptB, DptC, DptD, DptE, DptF and/or DptH, or all or part of a BAC clone described above. A host cell may comprise all or a part of an NRPS or PKS, such as a daptomycin NRPS. The host cell may further comprise one or more thioesterases.

Another object of the invention is to provide a polypeptide derived from the daptomycin biosynthetic gene cluster, preferably a polypeptide from the daptomycin biosynthetic gene cluster of *S. roseosporus*. The polypeptide may be DptA, DptB, DptC or DptD.

The invention also provides a polypeptide derived from an integral or free thioesterase, preferably one derived from a daptomycin biosynthetic gene cluster of *S. roseosporus*. In a preferred embodiment, the polypeptide is derived from thioesterase. The polypeptide may be derived from DptH or the thioesterase domain of DptD.

The invention also provides a polypeptide encoded by a nucleic acid molecule of any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05.

These polypeptides include, among others, enzymes involved in attaching a lipid tail to the peptide domain of daptomycin, polypeptides that regulate antibiotic resistance and ABC transporters.

Another object of the invention is to provide fragments of the polypeptides described above. In one embodiment, the fragment comprises at least one domain or module, as defined herein. In another embodiment, the fragment comprises at least one epitope of the polypeptide.

Another object of the invention is to provide polypeptides that are mutant proteins, fusion proteins, homologous proteins or allelic variants of the daptomycin NRPS polypeptides, thioesterases and polypeptides encoded by the nucleic acid molecules of the BAC clones provided herein.

The invention also provides an antibody that specifically binds to a polypeptide of a daptomycin NRPS, a thioesterase polypeptide of a daptomycin biosynthetic gene cluster or a polypeptide encoded by a nucleic acid molecule from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05. The invention also provides an antibody that can bind to a fragment, polypeptide mutant, a fusion protein, a polypeptide encoded by an allelic variant or a homologous protein of any one of the above-described polypeptides or proteins. The antibodies may be used to detect the presence or amount of a polypeptide of the instant invention or to inhibit or activate an activity of a polypeptide.

Another objective of the instant invention is to provide a method for recombinantly producing a polypeptide using a nucleic acid molecule described herein by introducing a nucleic acid molecule into a host cell and expressing the polypeptide.

The instant invention also provides a method for using the nucleic acid molecules of the instant invention to detect or amplify nucleic acid molecules that have similar or identical nucleic acid sequences compared to the nucleic acid molecules described herein.

The nucleic acid molecules and polypeptides are useful for, for example, the biosynthesis and production of natural products and the engineered biosynthesis of new compounds. The daptomycin NRPS and/or thioesterases may be used to produce daptomycin and other lipopeptides, including both naturally-occurring and novel compounds. The polypeptides may be used *in vitro* for the production of cyclic or non-cyclic lipopeptides, as well as other compounds produced by non-ribosomal peptide synthesis. Alternatively, a nucleic acid molecule of the invention may be introduced and expressed in a host cell, and the host cell may then be used to produce lipopeptides and other compounds produced by non-ribosomal peptide synthesis.

Another objective of the invention is to provide a novel gene cluster that can produce novel compounds by non-ribosomal peptide synthesis. A novel gene cluster may be obtained by altering nucleotides of the daptomycin biosynthetic gene cluster, particularly by altering nucleotides, domains or modules of the daptomycin NRPS, to make new polypeptides that are involved in non-ribosomal peptide synthesis. In this manner, different amino acids may be incorporated into a peptide produced by non-ribosomal peptide synthesis than the peptide produced by a naturally-occurring polypeptide. The invention also encompasses the compounds produced by the methods described herein.

Another objective of the invention is to provide a computer readable means of storing the nucleic acid and amino acid sequences of the instant invention. The records of the computer readable means can be accessed for reading and display of sequences and for comparison, alignment and ordering of the sequences of the invention to other sequences.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic diagram of methods in which daptomycin NRPS genes can be manipulated to alter gene expression or expression of the encoded proteins.

Figure 2A is a schematic diagram of BAC clone B12:03A05. The diagram shows a 90 kb region, referred to as the 90 kb fragment, and an approximately 12 kb region, referred to herein as the SP6 fragment. SEQ ID NO: 1 shows the nucleic acid sequence of the 90 kb fragment. SEQ ID NO: 103 shows the nucleic acid sequence of the SP6 fragment. The SP6 fragment abuts the 90 kb fragment. There is approximately 25-28 kb to the right of the 90 kb fragment (the GTC fragment).

Figure 2B shows a schematic diagram of the 90 kb fragment. There are 38 open reading frames (ORFs), which are nucleic acid sequences that encode polypeptides, in the region of the daptomycin biosynthetic gene cluster.

Figure 2C shows a schematic diagram of the SP6 fragment. There are 9 ORFs in the SP6 fragment. See Table 5 for the amino acid and nucleic acid sequence identifiers for the ORFs of the 90 kb and the SP6 fragment.

Figure 3 shows a comparison of the amino acid sequences of DptD (SEQ ID NO: 7) and the CDA III protein of *Streptomyces coelicolor* (SEQ ID NO:) using the Clustal W program. See Example 3.

Figure 4 shows a comparison of the amino acid sequences of DptH (SEQ ID NO: 8) and the CDA III protein of *Streptomyces coelicolor* using the Clustal W program. See Example 3.

Figures 5A-5C shows an analysis of daptomycin produced from the *Streptomyces lividans* TK64 clone containing the daptomycin biosynthetic gene cluster. Figure 5A shows an HPLC analysis of the broth of *Streptomyces lividans* TK64 clone containing BAC clone B12:03A05. The lower panel shows a trace plotting the maximum absorbance observed over the range of 200-600 nm for the HPLC eluate against time. The presence of three native lipopeptides, lipopeptides A21978C1 (the C1 lipopeptide), A21978C2 (the C2 lipopeptide) and A21978C3 (the C3 lipopeptide), is indicated by peaks with retention times of 5.61, 5.77 and 5.89 minutes, respectively. The upper panel shows the UV-visible spectra observed for these peaks. Figure 5B shows an ESI mass spectrum of daptomycin purified from decanoic acid-fed fermentation of *Streptomyces lividans* TK64 clone containing the daptomycin gene cluster. Figure 5C shows a 1H NMR spectrum (400MHz, in d6-DMSO) of daptomycin purified from decanoic acid-fed fermentation of *Streptomyces lividans* TK64 clone containing the daptomycin gene cluster.

Figure 6 is a diagram of the cloning vector pStreptoBAC V.

Figure 7 shows a *Hin*DIII digest of BAC clones from the Daptomycin biosynthetic gene cluster. Lane 1 shows 01G05 (82 kb insert); Lane 2 shows 03A05 (120 kb insert); Lane 3 shows 06A12 (85 kb insert); Lane 3 shows 12FG06 (65 kb insert); Lane 5 shows 18H04 (46 kb insert) and Lane 6 shows 20C09 (65 kb insert).

Figure 8 shows a map of some BAC clones that cover approximately 180 to 200 kb of the daptomycin NPRS region in *Streptomyces roseosporus*.

Figure 9 is a schematic diagram of the gene structure of an NRPS.

Figure 10 is a dendrogram showing the adenylation (A) domain similarities for domains that specify Asn and Asp in the daptomycin NRPS and in the Cda NRPS from *Streptomyces coelicolor*. See Example 5.

Figure 11 shows the results of an HPLC analysis determining the stereochemistry of Asn. See Example 6.

Figure 12 is a schematic diagram showing the organization of the daptomycin NRPS.

DETAILED DESCRIPTION OF THE INVENTION

Definitions and General Techniques

Unless otherwise defined herein, scientific and technical terms used in connection with the present invention shall have the meanings that are commonly understood by those of ordinary skill in the art. Further, unless otherwise required by context, singular terms shall include pluralities and plural terms shall include the singular. Generally, nomenclatures used in connection with, and techniques of, cell and tissue culture, molecular biology, immunology, microbiology, genetics and protein and nucleic acid chemistry and hybridization described herein are those well known and commonly used in the art. The methods and techniques of the present invention are generally performed according to conventional methods well known in the art and as described in various general and more specific references that are cited and discussed throughout the present specification unless otherwise indicated. *See, e.g.*, Sambrook et al. *Molecular Cloning: A Laboratory Manual,* 2d ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1989); Sambrook et al. *Molecular Cloning: A Laboratory Manual,* 3d ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (2000); Ausubel et al., *Current Protocols in Molecular Biology*, Greene Publishing Associates (1992, and Supplements to 2000); Ausubel et al., *Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology*, 4th ed., Wiley & Sons (1999); Harlow and Lane *Antibodies: A*

Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1990); Harlow and Lane *Using Antibodies: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1998); and T. Kieser et al., *Practical Streptomyces Genetics*, John Innes Foundation, Norwich (2000); each of which is incorporated herein by reference in its entirety.

Enzymatic reactions and purification techniques are performed according to manufacturer's specifications, as commonly accomplished in the art or as described herein. The nomenclatures used in connection with, and the laboratory procedures and techniques of, analytical chemistry, synthetic organic chemistry, and medicinal and pharmaceutical chemistry described herein are those well known and commonly used in the art. Standard techniques are used for chemical syntheses, chemical analyses, pharmaceutical preparation, formulation, and delivery, and treatment of patients.

The following terms, unless otherwise indicated, shall be understood to have the following meanings:

The term "thioesterase" refers to an enzyme that is capable of catalyzing the cleavage of a thioester bond, which may result in the production of a cyclic or linear molecule.

The term "thioesterase activity" refers to an enzymatic activity of a thioesterase, or a mutein, homologous protein, analog, derivative, fusion protein or fragment thereof, that catalyzes cleavage of a thioester bond. A thioesterase activity includes, e.g., an association and/or dissociation constants, a catalytic rate and a substrate turnover rate. A thioesterase activity of a polypeptide may be the same as one of the thioesterase activities of DptH, the thioesterase domain of DptD, a polypeptide encoded by *dptH*, a polypeptide encoded by the thioesterase domain of *dptD*, a polypeptide having an amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or a polypeptide having the amino acid sequence of SEQ ID NO: 8. The thioesterase activity may also different from that of one of the above-described thioesterases; e.g., it may have an increased or decreased catalytic activity, a different association and/or dissociation constant or a different substrate for catalysis. A "decreased" or "increased" thioesterase activity refers to a decreased or increased catalytic activity of the thioesterase, respectively.

A “thioesterase derived from a daptomycin biosynthetic gene cluster” is a thioesterase or thioesterase domain that is encoded by one of the genes of a gene cluster that encodes polypeptides involved in the synthesis of daptomycin. Preferably, the thioesterase is derived from a daptomycin biosynthetic gene cluster from *Streptomyces*, preferably from a daptomycin biosynthetic gene cluster from *S. roseosporus*.

A "daptomycin biosynthetic gene cluster" is defined herein as a nucleic acid molecule that encodes a number of polypeptides that are necessary for synthesis of daptomycin in an organism, preferably in a bacterial cell. A daptomycin biosynthetic gene cluster comprises a nucleic acid molecule that encodes at least DptA, DptB, DptC, DptD and DptH, or that encode muteins, homologous proteins, allelic variants or fragments thereof, as well as other nucleic acid sequences that encode other polypeptides required for daptomycin synthesis. Preferably, a daptomycin biosynthetic gene cluster comprises that part of BAC B12:03A05 that permits the synthesis of daptomycin when the part is introduced and expressed in a bacterial cell.

A "daptomycin NRPS" is defined herein as an NRPS that is capable of synthesizing daptomycin in an appropriate bacterial cell. A daptomycin NRPS comprises polypeptide subunits DptA, DptB, DptC and DptD, or muteins, homologous proteins, allelic variants or fragments thereof, that are capable, when expressed in an appropriate cell, of directing the synthesis of daptomycin. A daptomycin NRPS may further comprise DptH and/or other polypeptide, such as DptE or DptF. Preferably, the daptomycin NRPS is derived from the daptomycin biosynthetic gene cluster from *Streptomyces*, more preferably, the daptomycin NRPS is derived from *S. roseosporus*. The term "daptomycin NRPS" does not imply that the daptomycin NRPS can be used to synthesize only daptomycin. Rather, as used herein, the term is used solely for the purpose of describing that the NRPS was originally derived from a daptomycin biosynthetic gene cluster. The daptomycin NRPS may be used to synthesize molecules other than daptomycin, as described herein.

A “gene” is defined as a nucleic acid molecule that comprises a nucleic acid sequence that encodes a polypeptide and the expression control sequences that are operably linked to the nucleic acid sequence that encodes the polypeptide. For instance, a gene may comprise a promoter, one or more enhancers, a nucleic acid sequence that encodes a polypeptide, downstream regulatory sequences and, possibly, other nucleic acid sequences involved in regulation of the expression of an RNA.

A nucleic acid molecule or polypeptide is "derived" from a particular species if the nucleic acid molecule or polypeptide has been isolated from the particular species, or if the nucleic acid molecule or polypeptide is homologous to a nucleic acid molecule or polypeptide isolated from a particular species.

The terms "*dptA*", "*dptB*", "*dptC*" and "*dptD*" refer to nucleic acid molecules that encode subunits of the daptomycin NRPS. In a preferred embodiment, the nucleic acid molecule is derived from *Streptomyces*, more preferably the nucleic acid molecule is derived from *S. roseosporus*. In a preferred embodiment, the *dptA*, *dptB, dptC* and *dptD* encode the polypeptides having the amino acid sequences of SEQ ID NOS: 9, 11, 13 and 7, respectively. The terms "*dptA*", "*dptB*", "*dptC*" and "*dptD*" also refer to allelic variants of these genes, which may be obtained from other species of *Streptomyces* or from other *S. roseosporus* strains.

The term "*dptH*" refers to a gene whose coding domain encodes a thioesterase from a daptomycin biosynthetic gene cluster of *S. roseosporus*, wherein the naturally-occurring thioesterase is a "free" thioesterase. A free thioesterase is one that is not a functional domain of a larger polypeptide when it is naturally occurring. The *dptH* gene also encompasses the expression control sequences that are upstream of the coding region of the gene, as discussed below. In one embodiment, the expression control sequences of *dptH* have the nucleic acid sequence of SEQ ID NO: 5. The term "*dptH*" also refers to the nucleic acid encoding the polypeptide defined by SEQ ID NO: 8. The term "*dptH*" also refers to allelic variants of this gene, which may be obtained from other species of *Streptomyces* or from other *S. roseosporus* strains.

The term "allelic variant" refers to one of two or more alternative naturally-occurring forms of a gene, wherein each allele possesses a different nucleotide sequence. An allelic variant may encode the same polypeptide or a different one. As used herein, an allele is one that has at least 90% sequence identity, more preferably at least 95%, 96%, 97%, 98% or 99% sequence identity to the reference nucleic acid sequence, and encodes a polypeptide having similar or identical biological properties as the polypeptide encoded by the reference nucleic acid molecule.

The term “polynucleotide” or “nucleic acid molecule” refers to a polymeric form of nucleotides of at least 10 bases in length, either ribonucleotides or deoxynucleotides or a modified form of either type of nucleotide. The term includes single and double stranded forms of DNA. In addition, a polynucleotide may include either or both naturally-occurring and modified nucleotides linked together by naturally-occurring and/or non-naturally occurring nucleotide linkages.

An “isolated” or “substantially pure” nucleic acid or polynucleotide (e.g., an RNA, DNA or a mixed polymer) is one which is substantially separated from other cellular components that naturally accompany the native polynucleotide in its natural host cell, e.g., ribosomes, polymerases, or genomic sequences with which it is naturally associated. The term embraces a nucleic acid or polynucleotide that (1) has been removed from its naturally occurring environment, (2) is not associated with all or a portion of a polynucleotide in which the "isolated polynucleotide" is found in nature, (3) is operatively linked to a polynucleotide which it is not linked to in nature, or (4) does not occur in nature as part of a larger sequence. The term “isolated” or “substantially pure” also can be used in reference to recombinant or cloned DNA isolates, chemically synthesized polynucleotide analogs, or polynucleotide analogs that are biologically synthesized by heterologous systems.

A “part” of a nucleic acid molecule or polynucleotide refers to a nucleic acid molecule that comprises a partial contiguous sequence of at least 14 nucleotides of the reference nucleic acid molecule. Preferably, a part comprises at least 17 or 20 nucleotides of a reference nucleic acid molecule. More preferably, a part comprises at least 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 200, 300 400, 500 or 1000 nucleotides up to one nucleotide short of a reference nucleic acid molecule. A part of a nucleic acid molecule may comprise no other nucleic acid sequences. Alternatively, a part of a nucleic acid may comprise other nucleic acid sequences from other nucleic acid molecules.

The term "oligonucleotide" refers to a polynucleotide generally comprising a length of 200 nucleotides or fewer. Preferably, oligonucleotides are 10 to 60 nucleotides in length and most preferably 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50 or 60 nucleotides in length. Oligonucleotides may be single-stranded, e.g. for use as probes or primers, or may be double-stranded, e.g. for use in the construction of a mutant gene. Oligonucleotides of the invention can be either sense or antisense oligonucleotides. An oligonucleotide can include a label for detection, if desired.

The term "naturally-occurring nucleotide" referred to herein includes naturally-occurring deoxyribonucleotides and ribonucleotides. The term "modified nucleotides" referred to herein includes nucleotides with modified or substituted sugar groups and the like. The term "nucleotide linkages" referred to herein includes nucleotides linkages such as phosphorothioate, phosphorodithioate, phosphoroselenoate, phosphorodiselenoate, phosphoroanilothioate, phoshoraniladate, phosphoroamidate, and the like. *See e.g.*, LaPlanche et al. *Nucl. Acids Res.* 14:9081 (1986); Stec et al. *J. Am. Chem. Soc.* 106:6077 (1984); Stein et al. *Nucl. Acids Res.* 16:3209 (1988); Zon et al. *Anti-Cancer Drug Design* 6:539 (1991); Zon et al. *Oligonucleotides and Analogues: A Practical Approach*, pp. 87-108 (F. Eckstein, Ed., Oxford University Press, Oxford England (1991)); Stec et al. U.S. Patent No. 5,151,510; Uhlmann and Peyman *Chemical Reviews* 90:543 (1990), the disclosures of which are hereby incorporated by reference.

Unless specified otherwise, the left hand end of a polynucleotide sequence in sense orientation is the 5' end and the right hand end of the sequence is the 3' end. In addition, the left hand direction of a polynucleotide sequence in sense orientation is referred to as the 5' direction, while the right hand direction of the polynucleotide sequence is referred to as the 3' direction.

The term "percent sequence identity" or "identical" in the context of nucleic acid sequences refers to the residues in the two sequences which are the same when aligned for maximum correspondence. The length of sequence identity comparison may be over a stretch of at least about nine nucleotides, usually at least about 20 nucleotides, more usually at least about 24 nucleotides, typically at least about 28 nucleotides, more typically at least about 32 nucleotides, and preferably at least about 36 or more nucleotides. There are a number of different algorithms known in the art which can be used to measure nucleotide sequence identity. In one embodiment, polynucleotide sequences may be compared using Blast (Altschul et al., J. Mol. Biol. 215: 403-410, 1990). For instance, polynucleotide sequences can be compared using FASTA, Gap or Bestfit, which are programs in Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, Wisconsin. FASTA provides alignments and percent sequence identity of the regions of the best overlap between the query and search sequences (Pearson, 1990, (herein incorporated by reference). For instance, percent sequence identity between nucleic acid sequences can be determined using FASTA with its default parameters (a word size of 6 and the NOPAM factor for the scoring matrix) or using Gap with its default parameters as provided in GCG Version 6.1, herein incorporated by reference.

The term "substantial homology" or "substantial similarity," when referring to a nucleic acid or fragment thereof, indicates that, when optimally aligned with appropriate nucleotide insertions or deletions with another nucleic acid (or its complementary strand), there is nucleotide sequence identity in at least about 50%, more preferably 60% of the nucleotide bases, usually at least about 70%, more usually at least about 80%, preferably at least about 90%, and more preferably at least about 95%, 96%, 97%, 98% or 99% of the nucleotide bases, as measured by any well-known algorithm of sequence identity, such as FASTA, BLAST or Gap, as discussed above.

Alternatively, substantial homology or similarity exists when a nucleic acid or fragment thereof hybridizes to another nucleic acid, to a strand of another nucleic acid, or to the complementary strand thereof, under selective hybridization conditions. Typically, selective hybridization will occur when there is at least about 55% sequence identity -- preferably at least about 65%, more preferably at least about 75%, and most preferably at least about 90% -- over a stretch of at least about 14 nucleotides. See, e.g., Kanehisa, 1984, herein incorporated by reference.

Nucleic acid hybridization will be affected by such conditions as salt concentration, temperature, solvents, the base composition of the hybridizing species, length of the complementary regions, and the number of nucleotide base mismatches between the hybridizing nucleic acids, as will be readily appreciated by those skilled in the art. "Stringent hybridization conditions" and "stringent wash conditions" in the context of nucleic acid hybridization experiments depend upon a number of different physical parameters. The most important parameters include temperature of hybridization, base composition of the nucleic acids, salt concentration and length of the nucleic acid. One having ordinary skill in the art knows how to vary these parameters to achieve a particular stringency of hybridization.

In general, "stringent hybridization" is performed at about 25°C below the thermal melting point (T_m) for the specific DNA hybrid under a particular set of conditions. "Stringent washing" is performed at temperatures about 5°C lower than the T_m for the specific DNA hybrid under a particular set of conditions. The T_m is the temperature at which 50% of the target sequence hybridizes to a perfectly matched probe. See Sambrook et al., *supra*, page 9.51, hereby incorporated by reference.

The T_m for a particular DNA-DNA hybrid can be estimated by the formula:

$T_m = 81.5°C + 16.6 (\log_{10}[Na^+]) + 0.41$ (fraction $G + C$) - 0.63 (% formamide) - (600/l) where l is the length of the hybrid in base pairs.

The T_m for a particular RNA-RNA hybrid can be estimated by the formula:

$T_m = 79.8°C + 18.5 (\log_{10}[Na^+]) + 0.58$ (fraction $G + C$) + 11.8 (fraction $G + C)^2$ - 0.35 (% formamide) - (820/l).

The T_m for a particular RNA-DNA hybrid can be estimated by the formula:

$T_m = 79.8°C + 18.5(\log_{10}[Na^+]) + 0.58$ (fraction $G + C$) + 11.8 (fraction $G + C)^2$ - 0.50 (% formamide) - (820/l).

In general, the T_m decreases by 1-1.5°C for each 1% of mismatch between two nucleic acid sequences. Thus, one having ordinary skill in the art can alter hybridization and/or washing conditions to obtain sequences that have higher or lower degrees of sequence identity to the target nucleic acid. For instance, to obtain hybridizing nucleic acids that contain up to 10% mismatch from the target nucleic acid sequence, 10-15°C would be subtracted from the calculated T_m of a perfectly matched hybrid, and then the hybridization and washing temperatures adjusted accordingly. Probe sequences may also hybridize specifically to duplex DNA under certain conditions to form triplex or other higher order DNA complexes. The preparation of such probes and suitable hybridization conditions are well known in the art.

An example of stringent hybridization conditions for hybridization of complementary nucleic acid sequences having more than 100 complementary residues on a filter in a Southern or Northern blot or for screening a library is 50% formamide/6X SSC at 42°C for at least ten hours, preferably 12-16 hours. Another example of stringent hybridization conditions is 6X SSC at 68°C without formamide for at least ten hours, preferably 12-16 hours. An example of low stringency hybridization conditions for hybridization of complementary nucleic acid sequences having more than 100 complementary residues on a filter in a Southern or northern blot or for screening a library is 6X SSC at 42°C for at least ten hours, preferably 12-16 hours. Hybridization conditions to identify nucleic acid sequences that are similar but not identical can be identified by experimentally changing the hybridization temperature from 68°C to 42°C while keeping the salt concentration constant (6X SSC), or keeping the hybridization temperature and salt concentration constant (e.g. 42°C and 6X SSC) and varying the formamide concentration from 50% to 0%. Hybridization buffers may also include blocking agents to lower background. These agents are well-known in the art. See Sambrook et al., *supra*, pages 8.46 and 9.46-9.58, herein incorporated by reference.

Wash conditions also can be altered to change stringency conditions. An example of stringent wash conditions is a 0.2x SSC wash at 65°C for 15 minutes (see Sambrook et al., *supra*, for SSC buffer). Often the high stringency wash is preceded by a low stringency wash to remove excess probe. An exemplary medium stringency wash for duplex DNA of more than 100 base pairs is 1x SSC at 45°C for 15 minutes. An exemplary low stringency wash for such a duplex is 4x SSC at 40°C for 15 minutes. In general, signal-to-noise ratio of 2x or higher than that observed for an unrelated probe in the particular hybridization assay indicates detection of a specific hybridization.

As defined herein, nucleic acids that do not hybridize to each other under stringent conditions are still substantially homologous to one another if they encode polypeptides that are substantially identical to each other. This occurs, for example, when a nucleic acid is created synthetically or recombinantly using a high codon degeneracy as permitted by the redundancy of the genetic code.

The polynucleotides of this invention may include both sense and antisense strands of RNA, cDNA, genomic DNA, and synthetic forms and mixed polymers of the above. They may be modified chemically or biochemically or may contain non-natural or derivatized nucleotide bases, as will be readily appreciated by those of skill in the art. Such modifications include, for example, labels, methylation, substitution of one or more of the naturally occurring nucleotides with an analog, internucleotide modifications such as uncharged linkages (e.g., methyl phosphonates, phosphotriesters, phosphoramidates, carbamates, etc.), charged linkages (e.g., phosphorothioates, phosphorodithioates, etc.), pendent moieties (e.g., polypeptides), intercalators (e.g., acridine, psoralen, etc.), chelators, alkylators, and modified linkages (e.g., alpha anomeric nucleic acids, etc.) Also included are synthetic molecules that mimic polynucleotides in their ability to bind to a designated sequence via hydrogen bonding and other chemical interactions. Such molecules are known in the art and include, for example, those in which peptide linkages substitute for phosphate linkages in the backbone of the molecule.

The term "mutated" when applied to nucleic acid sequences means that nucleotides in a nucleic acid sequence may be inserted, deleted or changed compared to a reference nucleic acid sequence. A single alteration may be made at a locus (a point mutation) or multiple nucleotides may be inserted, deleted or changed at a single locus. In addition, one or more alterations may be made at any number of loci within a nucleic acid sequence. In a preferred embodiment, the nucleic acid sequence is the wild type nucleic acid sequence for a thioesterase. The nucleic acid sequence may be mutated by any method known in the art including those mutagenesis techniques described *infra*.

The term "error-prone PCR" refers to a process for performing PCR under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. See, e.g., Leung, D. W., et al., Technique, 1, pp.11-15 (1989) and Caldwell, R. C. & Joyce G. F., PCR Methods Applic., 2, pp. 28-33 (1992).

The term "oligonucleotide-directed mutagenesis" refers to a process which enables the generation of site-specific mutations in any cloned DNA segment of interest. See, e.g., Reidhaar-Olson, J. F. & Sauer, R. T., et al., Science, 241, pp. 53-57 (1988).

The term "assembly PCR" refers to a process which involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions occur in parallel in the same vial, with the products of one reaction priming the products of another reaction.

The term "sexual PCR mutagenesis" of "DNA shuffling"refers to a method of error-prone PCR coupled with forced homologous recombination between DNA molecules of different but highly related DNA sequence *in vitro*, caused by random fragmentation of the DNA molecule based on sequence homology, followed by fixation of the crossover by primer extension in an error-prone PCR reaction. See, e.g., Stemmer, W. P., Proc. Natl. Acad. Sci. U.S.A., 91, pp. 10747-10751 (1994). DNA shuffling can be carried out between several related genes ("Family shuffling").

The term "*in vivo* mutagenesis" refers to a process of generating random mutations in any cloned DNA of interest which involves the propagation of the DNA in a strain of bacteria such as *E. coli* that carries mutations in one or more of the DNA repair pathways. These "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in a mutator strain will eventually generate random mutations within the DNA.

The term "cassette mutagenesis" refers to any process for replacing a small region of a double-stranded DNA molecule with a synthetic oligonucleotide "cassette" that differs from the native sequence. The oligonucleotide often contains completely and/or partially randomized native sequence.

The term "recursive ensemble mutagenesis" refers to an algorithm for protein engineering (protein mutagenesis) developed to produce diverse populations of phenotypically related mutants whose members differ in amino acid sequence. This method uses a feedback mechanism to control successive rounds of combinatorial cassette mutagenesis. See, e.g., Arkin, A. P. and Youvan, D. C., Proc. Natl. Acad. Sci. U.S.A., 89, pp. 7811-7815 (1992).

The term "exponential ensemble mutagenesis" refers to a process for generating combinatorial libraries with a high percentage of unique and functional mutants, wherein small groups of residues are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. See, e.g., Delegrave, S. and Youvan, D. C., Biotechnology Research, 11, pp. 1548-1552 (1993); and random and site-directed mutagenesis, Arnold, F. H., Current Opinion in Biotechnology, 4, pp. 450-455 (1993). Each of the references mentioned above are hereby incorporated by reference in its entirety.

"Operatively linked" expression control sequences refers to a linkage in which the expression control sequence is contiguous with the gene of interest to control the gene of interest, as well as expression control sequences that act in *trans* or at a distance to control the gene of interest.

The term "expression control sequence" as used herein refers to polynucleotide sequences which are necessary to affect the expression of coding sequences to which they are operatively linked. Expression control sequences are sequences which control the transcription, post-transcriptional events and translation of nucleic acid sequences. Expression control sequences include appropriate transcription initiation, termination, promoter and enhancer sequences; efficient RNA processing signals such as splicing and polyadenylation signals; sequences that stabilize cytoplasmic mRNA; sequences that enhance translation efficiency (e.g., ribosome binding sites); sequences that enhance protein stability; and when desired, sequences that enhance protein secretion. The nature of such control sequences differs depending upon the host organism; in prokaryotes, such control sequences generally include promoter, ribosomal binding site, and transcription termination sequence. The term "control sequences" is intended to include, at a minimum, all components whose presence is essential for expression, and can also include additional components whose presence is advantageous, for example, leader sequences and fusion partner sequences.

The term "vector," as used herein, is intended to refer to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. One type of vector is a "plasmid", which refers to a circular double stranded DNA loop into which additional DNA segments may be ligated. Other vectors include cosmids, bacterial artificial chromosomes (BAC) and yeast artificial chromosomes (YAC). Another type of vector is a viral vector, wherein additional DNA segments may be ligated into the viral genome. Viral vectors that infect bacterial cells are referred to as bacteriophages. Certain vectors are capable of autonomous replication in a host cell into which they are introduced (e.g., bacterial vectors having a bacterial origin of replication). Other vectors can be integrated into the genome of a host cell upon introduction into the host cell, and thereby are replicated along with the host genome. Moreover, certain vectors are capable of directing the expression of genes to which they are operatively linked. Such vectors are referred to herein as "recombinant expression vectors" (or simply, "expression vectors"). In general, expression vectors of utility in recombinant DNA techniques are often in the form of plasmids. In the present specification, "plasmid" and "vector" may be used interchangeably as the plasmid is the most commonly used form of vector. However, the invention is intended to include other forms of expression vectors that serve equivalent functions.

The term “recombinant host cell” (or simply "host cell"), as used herein, is intended to refer to a cell into which a recombinant expression vector has been introduced. It should be understood that such terms are intended to refer not only to the particular subject cell but to the progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term "host cell" as used herein.

The term "polypeptide" encompasses both naturally-occurring and non-naturally-occurring proteins and polypeptides, polypeptide fragments and polypeptide mutants, derivatives and analogs. As used herein, a polypeptide comprises at least six amino acids, preferably at least 8, 10, 12, 15, 20, 25 or 30 amino acids, and more preferably the polypeptide is the full length of the naturally-occurring polypeptide. A polypeptide may be monomeric or polymeric. Further, a polypeptide may comprise a number of different modules within a single polypeptide each of which has one or more distinct activities. A preferred polypeptide in accordance with the invention comprises a thioesterase derived from the daptomycin biosynthetic gene cluster, as well as a fragment, mutant, analog and derivative thereof.

The term "isolated protein" or “isolated polypeptide” is a protein or polypeptide that by virtue of its origin or source of derivation (1) is not associated with naturally associated components that accompany it in its native state, (2) is free of other proteins from the same species (3) is expressed by a cell from a different species, or (4) does not occur in nature. Thus, a polypeptide that is chemically synthesized or synthesized in a cellular system different from the cell from which it naturally originates will be "isolated" from its naturally associated components. A polypeptide or protein may also be rendered substantially free of naturally associated components by isolation, using protein purification techniques well known in the art.

A protein or polypeptide is "substantially pure," "substantially homogeneous" or "substantially purified" when at least about 60% to 75% of a sample exhibits a single species of polypeptide. The polypeptide or protein may be monomeric or multimeric. A substantially pure polypeptide or protein will typically comprise about 50%, 60%, 70%, 80% or 90% W/W of a protein sample, more usually about 95%, and preferably will be over 99% pure. Protein purity or homogeneity may be indicated by a number of means well known in the art, such as polyacrylamide gel electrophoresis of a protein sample, followed by visualizing a single polypeptide band upon staining the gel with a stain well known in the art. For certain purposes, higher resolution may be provided by using HPLC or other means well known in the art for purification.

The term "polypeptide fragment" as used herein refers to a polypeptide that has an amino-terminal and/or carboxy-terminal deletion compared to a full-length polypeptide. In a preferred embodiment, the polypeptide fragment is a contiguous sequence in which the amino acid sequence of the fragment is identical to the corresponding positions in the naturally-occurring sequence. Fragments typically are at least 6, 7, 8, 9 or 10 amino acids long, preferably at least 12, 14, 16 or 18 amino acids long, more preferably at least 20 amino acids long, more preferably at least 25, 30, 35, 40 or 45, amino acids, even more preferably at least 50 or 60 amino acids long, and even more preferably at least 70 amino acids long.

A "derivative" refers to polypeptides or fragments thereof that are substantially homologous in primary structural sequence but which include, e.g., *in vivo* or *in vitro* chemical and biochemical modifications or which incorporate amino acids that are not found in the native polypeptide. Such modifications include, for example, acetylation, carboxylation, phosphorylation, glycosylation, ubiquitination, labeling, e.g., with radionuclides, and various enzymatic modifications, as will be readily appreciated by those well skilled in the art. A variety of methods for labeling polypeptides and of substituents or labels useful for such purposes are well known in the art, and include radioactive isotopes such as ^{125}I, ^{32}P, ^{35}S, and ^{3}H, ligands which bind to labeled antiligands (e.g., antibodies), fluorophores, chemiluminescent agents, enzymes, and antiligands which can serve as specific binding pair members for a labeled ligand. The choice of label depends on the sensitivity required, ease of conjugation with the primer, stability requirements, and available instrumentation. Methods for labeling polypeptides are well known in the art. See Ausubel et al., 1992, hereby incorporated by reference.

The term "fusion protein" refers to polypeptides comprising polypeptides or fragments coupled to heterologous amino acid sequences. Fusion proteins are useful because they can be constructed to contain two or more desired functional elements from two or more different proteins. A fusion protein comprises at least 10 contiguous amino acids from a polypeptide of interest, more preferably at least 20 or 30 amino acids, even more preferably at least 40, 50 or 60 amino acids, yet more preferably at least 75, 100 or 125 amino acids. Fusion proteins can be produced recombinantly by constructing a nucleic acid sequence which encodes the polypeptide or a fragment thereof in frame with a nucleic acid sequence encoding a different protein or peptide and then expressing the fusion protein. Alternatively, a fusion protein can be produced chemically by crosslinking the polypeptide or a fragment thereof to another protein.

The term "non-peptide analog" refers to a compound with properties that are analogous to those of a reference polypeptide. A non-peptide compound may also be termed a "peptide mimetic" or a "peptidomimetic." See, e.g., Fauchere, *J. Adv. Drug Res.* 15:29 (1986); Veber and Freidinger *TINS* p.392 (1985); and Evans et al. *J. Med. Chem*. 30:1229 (1987), which are incorporated herein by reference. Such compounds are often developed with the aid of computerized molecular modeling. Peptide mimetics that are structurally similar to useful peptides may be used to produce an equivalent effect. Generally, peptidomimetics are structurally similar to a paradigm polypeptide (i.e., a polypeptide that has a desired biochemical property or pharmacological activity), such as a thioesterase, but have one or more peptide linkages optionally replaced by a linkage selected from the group consisting of: $--CH_2NH--$, $--CH_2S--$, $--CH_2-CH_2--$, $--CH=CH--$(cis and trans), $--COCH_2--$,

--CH(OH)CH_2--, and –CH_2SO--, by methods well known in the art. Systematic substitution of one or more amino acids of a consensus sequence with a D-amino acid of the same type (e.g., D-lysine in place of L-lysine) may also be used to generate more stable peptides. In addition, constrained peptides comprising a consensus sequence or a substantially identical consensus sequence variation may be generated by methods known in the art (Rizo and Gierasch *Ann. Rev. Biochem.* 61:387 (1992), incorporated herein by reference); for example, by adding internal cysteine residues capable of forming intramolecular disulfide bridges which cyclize the peptide.

A "polypeptide mutant" or "mutein" refers to a polypeptide whose sequence contains substitutions, insertions or deletions of one or more amino acids compared to the amino acid sequence of a native or wild type protein. A mutein may have one or more amino acid point substitutions, in which a single amino acid at a position has been changed to another amino acid, one or more insertions and/or deletions, in which one or more amino acids are inserted or deleted, respectively, in the sequence of the naturally-occurring protein, and/or truncations of the amino acid sequence at either or both the amino or carboxy termini. Further, a mutein may have the same or different biological activity as the naturally-occurring protein. For instance, a mutein may have an increased or decreased biological activity. In a preferred embodiment of the present invention, a mutein has the same or increased thioesterase activity as a naturally-occurring thioesterase. A mutein has at least 50%, 60% or 70% sequence homology to the wild type protein, more preferred are muteins having at least 80%, 85% or 90% sequence homology to the wild type protein, even more preferred are muteins exhibiting at least 95%, 96%, 97%, 98% or 99% sequence identity. Sequence homology may be measured by any common sequence analysis algorithm, such as Gap or Bestfit, using default parameters.

Preferred amino acid substitutions are those which: (1) reduce susceptibility to proteolysis, (2) reduce susceptibility to oxidation, (3) alter binding affinity for forming protein complexes, (4) alter binding affinity or enzymatic activity, and (5) confer or modify other physicochemical or functional properties of such derivatives, analogs, fusion proteins and muteins. Single or multiple amino acid substitutions (preferably conservative amino acid substitutions) may be made in the naturally-occurring sequence (preferably in the portion of the polypeptide outside the domain(s) forming intermolecular contacts. A conservative amino acid substitution should not substantially change the structural characteristics of the parent sequence (e.g., a replacement amino acid should not tend to break a helix that occurs in the parent sequence, or disrupt other types of secondary structure that characterizes the parent sequence). Examples of art-recognized polypeptide secondary and tertiary structures are described in *Proteins, Structures and Molecular Principles* (Creighton, Ed., W. H. Freeman and Company, New York (1984)); *Introduction to Protein Structure* (C. Branden and J. Tooze, eds., Garland Publishing, New York, N.Y. (1991)); and Thornton et at. *Nature* 354:105 (1991), which are each incorporated herein by reference.

As used herein, the twenty conventional amino acids and their abbreviations follow conventional usage. *See Immunology - A Synthesis* (2nd Edition, E.S. Golub and D.R. Gren, Eds., Sinauer Associates, Sunderland, Mass. (1991)), which is incorporated herein by reference. Stereoisomers (e.g., D-amino acids) of the twenty conventional amino acids, unnatural amino acids such as α-, α-disubstituted amino acids, N-alkyl amino acids, and other unconventional amino acids may also be suitable components for polypeptides of the present invention. Examples of unconventional amino acids include: 4-hydroxyproline, γ-carboxyglutamate, ϵ-N,N,N-trimethyllysine, ϵ-N-acetyllysine, O-phosphoserine, N-acetylserine, N-formylmethionine, 3-methylhistidine, 5-hydroxylysine, s-N-methylarginine, and other similar amino acids and imino acids (e.g., 4-hydroxyproline). In the polypeptide notation used herein, the lefthand direction is the amino terminal direction and the right hand direction is the carboxy-terminal direction, in accordance with standard usage and convention.

A protein has “homology” or is “homologous” to a protein from another organism if the encoded amino acid sequence of the protein has a similar sequence to the encoded amino acid sequence of a protein of a different organism. Alternatively, a protein may have homology or be homologous to another protein if the two proteins have similar amino acid sequences. Although two proteins are said to be “homologous,” this does not imply that there is necessarily an evolutionary relationship between the proteins. Instead, the term “homologous” is defined to mean that the two proteins have similar amino acid sequences. In a preferred embodiment, a homologous protein is one that exhibits at least 50%, 60% or 70% sequence identity to the wild type protein, preferred are homologous proteins that exhibit at least 80%, 85%, 90%, 95%, 96%, 97%, 98% or 99% sequence identity. In addition, although in many cases proteins with similar amino acid sequences will have similar functions, the term "homologous" does not imply that the proteins must be functionally similar to each other.

When "homologous" is used in reference to proteins or peptides, it is recognized that residue positions that are not identical often differ by conservative amino acid substitutions. A "conservative amino acid substitution" is one in which an amino acid residue is substituted by another amino acid residue having a side chain ® group) with similar chemical properties (e.g., charge or hydrophobicity). In general, a conservative amino acid substitution will not substantially change the functional properties of a protein. In cases where two or more amino acid sequences differ from each other by conservative substitutions, the percent sequence identity or degree of homology may be adjusted upwards to correct for the conservative nature of the substitution. Means for making this adjustment are well known to those of skill in the art (see, e.g., Pearson et al.,1994, herein incorporated by reference).

The following six groups each contain amino acids that are conservative substitutions for one another:

1) Serine (S), Threonine (T);
2) Aspartic Acid (D), Glutamic Acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Alanine (A), Valine (V), and
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W).

Sequence homology for polypeptides, which is also referred to as sequence identity, is typically measured using sequence analysis software. See, e.g., the Sequence Analysis Software Package of the Genetics Computer Group (GCG), University of Wisconsin Biotechnology Center, 910 University Avenue, Madison, Wisconsin 53705. Protein analysis software matches similar sequences using measure of homology assigned to various substitutions, deletions and other modifications, including conservative amino acid substitutions. For instance, GCG contains programs such as "Gap" and "Bestfit" which can be used with default parameters to determine sequence homology or sequence identity between closely related polypeptides, such as homologous polypeptides from different species of organisms or between a wild type protein and a mutein thereof. See, e.g., GCG Version 6.1.

A preferred algorithm when comparing a polypeptide sequence to a database containing a large number of sequences from different organisms is the computer program BLAST, especially blastp, tblastn or BlastX. See Altschul et al. Nucleic Acids Res. 25:3389-3402 (1997), herein incorporated by reference. BlastX, which compares a translated nucleotide sequence to a protein database, may be performed through the servers located at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov). Preferred parameters for blastp, which compares a protein sequence to a protein database are:

Expectation value:	10 (default)
Filter:	seg (default)
Cost to open a gap:	11 (default)
Cost to extend a gap:	1 (default
Max. alignments:	100 (default)
Word size:	11 (default)
No. of descriptions:	100 (default)
Penalty Matrix:	BLOSUM62

The length of polypeptide sequences compared for homology will generally be at least about 16 amino acid residues, usually at least about 20 residues, more usually at least about 24 residues, typically at least about 28 residues, and preferably more than about 35 residues. When searching a database containing sequences from a large number of different organisms, it is preferable to compare amino acid sequences.

Database searching using amino acid sequences can be measured by algorithms other than blastp known in the art. For instance, polypeptide sequences can be compared using FASTA, a program in GCG Version 6.1. FASTA provides alignments and percent sequence identity of the regions of the best overlap between the query and search sequences (Pearson, 1990, herein incorporated by reference). For example, percent sequence identity between amino acid sequences can be determined using FASTA with its default parameters (a word size of 2 and the PAM250 scoring matrix), as provided in GCG Version 6.1, herein incorporated by reference.

An "antibody" refers to an intact immunoglobulin, or to an antigen-binding portion thereof that competes with the intact antibody for antigen-specific binding. Antigen-binding portions may be produced by recombinant DNA techniques or by enzymatic or chemical cleavage of intact antibodies. Antigen-binding portions include, *inter alia*, Fab, Fab', $F(ab')_{2,}$ Fv, dAb, and complementarity determining region (CDR) fragments, single-chain antibodies (scFv), chimeric antibodies, diabodies and polypeptides that contain at least a portion of an immunoglobulin that is sufficient to confer specific antigen binding to the polypeptide. An Fab fragment is a monovalent fragment consisting of the VL, VH, CL and CH1 domains; a $F(ab')_2$ fragment is a bivalent fragment comprising two Fab fragments linked by a disulfide bridge at the hinge region; a Fd fragment consists of the VH and CH1 domains; an Fv fragment consists of the VL and VH domains of a single arm of an antibody; and a dAb fragment (Ward et al., Nature 341:544-546, 1989) consists of a VH domain.

A single-chain antibody (scFv) is an antibody in which a VL and VH regions are paired to form a monovalent molecules via a synthetic linker that enables them to be made as a single protein chain (Bird et al., Science 242:423-426, 1988 and Huston et al., Proc. Natl. Acad. Sci. USA 85:5879-5883, 1988). Diabodies are bivalent, bispecific antibodies in which VH and VL domains are expressed on a single polypeptide chain, but using a linker that is too short to allow for pairing between the two domains on the same chain, thereby forcing the domains to pair with complementary domains of another chain and creating two antigen binding sites (see e.g., Holliger, P., et al., Proc. Natl. Acad. Sci. USA 90:6444-6448, 1993, and Poljak, R. J., et al., Structure 2:1121-1123, 1994). One or more CDRs may be incorporated into a molecule either covalently or noncovalently to make it an immunoadhesin. An immunoadhesin may incorporate the CDR(s) as part of a larger polypeptide chain, may covalently link the CDR(s) to another polypeptide chain, or may incorporate the

CDR(s) noncovalently. The CDRs permit the immunoadhesin to specifically bind to a particular antigen of interest. A chimeric antibody is an antibody that contains one or more regions from one antibody and one or more regions from one or more other antibodies.

An antibody may have one or more binding sites. If there is more than one binding site, the binding sites may be identical to one another or may be different. For instance, a naturally-occurring immunoglobulin has two identical binding sites, a single-chain antibody or Fab fragment has one binding site, while a "bispecific" or "bifunctional" antibody has two different binding sites.

An “isolated antibody” is an antibody that (1) is not associated with naturally-associated components, including other naturally-associated antibodies, that accompany it in its native state, (2) is free of other proteins from the same species, (3) is expressed by a cell from a different species, or (4) does not occur in nature.

A “neutralizing antibody” or “an inhibitory antibody” is an antibody that inhibits the activity of a polypeptide or blocks the binding of a polypeptide to a ligand that normally binds to it. For example, a neutralizing anti-thioesterase antibody may be one that blocks the activity of the thioesterase. An “activating antibody” is an antibody that increases the activity of a polypeptide. For example, an activating anti-thioesterase antibody is one that increases the activity of a thioesterase.

The term "epitope" includes any protein determinant capable of specific binding to an immunoglobulin or T-cell receptor. Epitopic determinants usually consist of chemically active surface groupings of molecules such as amino acids or sugar side chains and usually have specific three dimensional structural characteristics, as well as specific charge characteristics. An antibody is said to specifically bind an antigen when the dissociation constant is ≤ 1 μM, preferably ≤ 100 nM and most preferably ≤ 10 nM.

The term patient includes human and veterinary subjects.

Throughout this specification and claims, the word "comprise," or variations such as "comprises" or "comprising," will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers.

Nucleic Acid Molecules, Regulatory Sequences, Vectors, Host Cells and Recombinant Methods of Making Polypeptides

Nucleic Acid Molecules

In one aspect, the present invention provides a nucleic acid molecule encoding a thioesterase or a daptomycin NRPS or a subunit thereof. In one embodiment, the nucleic acid molecule encodes one or more of DptA, DptB, DptC or DptD. In a preferred embodiment, the nucleic acid molecules encodes a polypeptide comprising any one of the amino acid sequences of SEQ ID NOS: 9, 11, 13 or 7. In another preferred embodiment, the nucleic acid molecule comprises *dptA, dptB, dptC* and/or *dptD*. In a further preferred embodiment, the nucleic acid molecule comprises a nucleic acid sequence comprising any one of SEQ ID NOS: 10, 12, 14 or 3.

In another embodiment, the nucleic acid molecule encodes a thioesterase that is derived from a daptomycin biosynthetic gene cluster. In a preferred embodiment, the nucleic acid molecule encodes a thioesterase derived from a daptomycin biosynthetic gene cluster that is a free thioesterase or is an integral thioesterase. In another preferred embodiment, the nucleic acid molecule encodes DptH or the thioesterase domain of DptD. In a more preferred embodiment, the nucleic acid molecule encodes a polypeptide comprising an amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or has the amino acid sequence of SEQ ID NO: 8. In another embodiment, the nucleic acid molecule comprises the thioesterase-encoding domain of *dptD* or *dptH* from the daptomycin biosynthetic gene cluster. In another preferred embodiment, the nucleic acid molecule comprises a nucleic acid sequence of SEQ ID NO: 6 or of SEQ ID NO: 3, or the region comprising the thioesterase-encoding portion thereof. In another embodiment, the nucleic acid molecule also encodes a daptomycin NRPS or a subunit thereof. See Examples 1-6 regarding the isolation and identification of *dptA, dptB, dptC, dptD* and *dptH* and other genes of the daptomycin biosynthetic gene cluster.

In another embodiment, the nucleic acid molecule encodes an acyl CoA ligase. In a preferred embodiment, the nucleic acid molecule encodes DptE, preferably a nucleic acid molecule encoding SEQ ID NO: 15. In a more preferred embodiment, the nucleic acid molecule comprises *dptE*. In an even more preferred embodiment, the nucleic acid molecule comprises SEQ ID NO: 16. In another embodiment, the nucleic acid molecule encodes an acyl transferase. In a preferred embodiment, the nucleic acid molecule encodes DptF, preferably a nucleic acid molecule encoding SEQ ID NO: 17. In a more preferred embodiment, the nucleic acid molecule comprises *dptF*. In an even more preferred embodiment, the nucleic acid molecule comprises SEQ ID NO: 18.

Another embodiment of the invention provides a nucleic acid molecule comprising a DNA sequence from a bacterial artificial chromosome (BAC) comprising nucleic acid sequences from *S. roseosporus*. In a preferred embodiment, the nucleic acid molecule comprises a *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05. In a preferred embodiment, the nucleic acid molecule comprises a *S. roseosporus* nucleic acid sequence from B12:03A05 (ATCC Deposit PTA-3140, deposited March 1, 2001). The nucleic acid molecule may comprise the entire *S. roseosporus* nucleic acid sequence in the BAC clone or may comprise a part thereof. In a preferred embodiment, the part is a nucleic acid molecule that comprises at least one nucleic acid sequence that can encode a polypeptide, preferably a full-length polypeptide, i.e., a nucleic acid molecule that encodes a polypeptide from its start codon to its stop codon. In one preferred embodiment, the part comprises a nucleic acid molecule encoding a polypeptide involved in daptomycin biosynthesis, such as, without limitation, *dptA, dptB, dptC, dptD, dptE, dptF* or *dptH*.

In another embodiment, a part from the BAC clone is a nucleic acid molecule comprising a nucleic acid sequence encoding a polypeptide selected from SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101. In another embodiment, the part from the BAC clone is a nucleic acid molecule comprising a nucleic acid sequence selected from SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. The polypeptides having amino acids sequences of SEQ ID NOS: 19, 21, 29, 45, 47, 49, 63, 67, 75 and 77 (nucleic acid sequences of SEQ ID NOS: 20, 22, 30, 46, 48, 50, 64, 68, 76 or 78) are ATP transporters. Some of the polypeptides are pump-like polypeptides with Walker motifs while others are polypeptides that have a role in metal scavenging, e.g., iron or manganese transport (see Tables 6 and 7). The nucleic acid molecule comprising SEQ ID NO: 76 encodes an ATP-binding component of an ABC transporter system, as determined by its sequence similarity to ORF1 of (AAD44229.1) of *S. rochei* and the *S. peucetius* DrrA (P32010) genes. The encoded polypeptide has both a Walker A and a Walker B motif. Further, its synthesis appears to be translationally coupled to that of a nucleic acid molecule comprising SEQ ID NO: 78, which encodes a DrrB-like polypeptide, as determined by its sequence similar to the *S. peuticeus* DrrB product (AAA74718.1), encoding the integral membrane component. The polypeptide having an amino acid sequence of SEQ ID NO: 21 is a *StrV* homolog, while the polypeptide having an amino acid sequence of SEQ ID NO: 19 is a *StrW* homolog. See, e.g., Beyer et al., 1996, *supra*. The *StrV* homolog has both Walker motifs, while the *StrW* homolog has only a Walker B motif. Both nucleic acid sequences encoding the polypeptide are on the complementary strand and appear to be translationally regulated. They have *S. coelicolor* homologs, G8A.01 and G8A.02 (emb| CAB88931, CAB88932). See Tables 6 and 7.

In another aspect, a part of the BAC clone is a nucleic acid molecule comprising a nucleic acid sequence encoding an oxidoreductase, a dehydrogenase; a transcriptional regulator involved in antibiotic resistance; NovABC-related polypeptides, which are involved in the biosynthesis of novobiocin, an antimicrobial agent; a monooxygenase; an acyl CoA thioesterase; a DNA helicase; or a DNA ligase. These nucleic acid molecules and encoded polypeptides may be useful in daptomycin biosynthesis; e.g., the acyl CoA thioesterase may be useful for the reasons provided above for thioesterases and may also be important in addition of the lipid tail to the peptide domain of daptomycin. These nucleic acid molecules encoding enzymes are also useful because they may be used in the same way as other oxidoreductases, dehydrogenases, monooxygenases, DNA helicases or DNA ligases are used in the art. Notably, the transcriptional regulator can be mutated using well-known methods to increase or decrease daptomycin or other antibiotic resistance. The nucleic acid molecules encoding NovABC-related polypeptides may be used in the same way as NovABC is used in the art, e.g., to produce novobiocin or related antimicrobial agents.

The polypeptides having the above-described activity comprise the amino acid sequences of SEQ ID NOS: 23, 25, 27, 29, 33, 35, 37, 91, 93, 97 and 99 and are encoded by nucleic acid sequences of SEQ ID NOS: 24, 26, 28, 30, 34, 36, 38, 92, 94, 98 and 100.

In another aspect, a part of the BAC clone is a nucleic acid molecule that encodes a polypeptide that does not have a defined function but which is highly homologous to nucleic acid molecules and polypeptides from other *Streptomyces*. These nucleic acid molecules (SEQ ID NOS: 62, 66, 70, 80, 82, 84, 86, 88, 96 and 102), the polypeptides they encode (SEQ ID NOS: 61, 65, 69, 79, 81, 83, 85, 87, 95 and 101) and antibodies to the polypeptides may be used to identify other *Streptomyces* species using standard molecular biological and protein chemistry techniques (e.g., PCR, RT-PCR, Southern blotting, northern blotting, ELISAs, radioimmunoassays or western blotting), which is useful, e.g., in microbiological testing or forensics. In another embodiment, a part of the BAC clone is a nucleic acid molecule that encodes a polypeptide that does not have a defined function and is not highly homologous to a nucleic acid molecule or polypeptide from another species. These nucleic acid molecules (SEQ ID NOS: 32, 40, 42, 44, 52, 54, 56, 58, 60, 72 and 74) are nevertheless useful because they are close to the daptomycin biosynthetic gene cluster, and as such, they can be used to identify nucleic acid molecules that encode all or a part of the daptomycin biosynthetic gene cluster. Parts of the BAC clone that do not encode a polypeptide are useful for the same reasons. Further, the polypeptides having the amino acid sequence of SEQ ID NOS: 31, 39, 41, 43, 51, 53, 55, 57, 59, 71 and 73 can be used to make antibodies that can be used to identify *S. roseosporus*. Because the polypeptides are not highly homologous to any other species, the antibodies would likely be highly specific for *S. roseosporus*.

In another aspect, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule as described above. In a preferred embodiment, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule that encodes DptA, DptB, DptC, DptD or DptH. In another preferred embodiment, the invention provides a nucleic acid molecules that selectively hybridizes to a nucleic acid molecule that encodes SEQ ID NOS: 9, 11, 13, 7 or 8. In an even more preferred embodiment, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule comprising the nucleic acid sequence of *dptA, dptB, dptC, dptD* or *dptH*. In another preferred embodiment, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule comprising the nucleic acid sequence SEQ ID NOS: 10, 12, 14, 3 or 6. The invention also provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule comprising an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably that from B12:03A05. In a preferred embodiment, the invention provides a nucleic acid molecule that selectively hybridizes to a nucleic acid molecule encoding SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or to a nucleic acid molecule comprising the nucleic acid sequence SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. The selective hybridization of any of the above-described nucleic acid sequences may be performed under low stringency hybridization conditions. In a preferred embodiment, the selective hybridization is performed under high stringency hybridization conditions. In a preferred embodiment of the invention, the hybridizing nucleic acid molecule may be used to recombinantly express a polypeptide of the invention.

In another aspect, the invention provides a nucleic acid molecule that is homologous to a nucleic acid encoding a daptomycin NRPS or subunit thereof, a thioesterase from a daptomycin biosynthetic gene cluster, or a nucleic acid molecule comprising an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or, preferably, B12:03A05. The invention provides a nucleic acid molecule homologous to a nucleic acid molecule encoding DptA, DptB, DptC, DptD or DptH. In one embodiment, the nucleic acid molecule is homologous to a nucleic acid molecule encoding a polypeptide having an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8. In a preferred embodiment, the nucleic acid molecule is homologous to any one or more of *dptA, dptB, dptC* or *dptD*. In another embodiment, the nucleic acid molecule is homologous to a thioesterase encoded by the thioesterase domain of *dptD* or by *dptH*. In a more preferred embodiment, the nucleic acid molecule is homologous to a nucleic acid molecule having a nucleic acid sequence of SEQ ID NOS: 10, 12, 14, 3 or 6. In another preferred embodiment, the invention provides a nucleic acid molecule that is homologous to a nucleic acid molecule encoding SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or to a nucleic acid molecule comprising the nucleic acid sequence SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. In a preferred embodiment, a homologous nucleic acid molecule is one that has at least 60%, 70%, 80% or 85% sequence identity with a nucleic acid molecule described herein. In a more preferred embodiment, the homologous nucleic acid molecule is one that has at least 90%, 95%, 97%, 98% or 99% sequence identity with a nucleic acid molecule described herein. Further, in one embodiment, a homologous nucleic acid molecule is homologous over its entire length to a nucleic acid molecule encoding a daptomycin NRPS or subunit thereof, a thioesterase, or nucleic acid molecule that encodes a polypeptide as described herein. In another embodiment, a homologous nucleic acid molecule is homologous over only a part of its length to a nucleic acid molecule described herein, wherein the part is at least 50 nucleotides of the nucleic acid molecule, preferably at least 100 nucleotides, more preferably at least 200 nucleotides, even more preferably at least 300 nucleotides.

In another embodiment, the invention provides a nucleic acid that is an allelic variant of a gene encoding a daptomycin NRPS or subunit thereof, a thioesterase from a daptomycin biosynthetic gene cluster, or a nucleic acid molecule comprising an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05. In a preferred embodiment, the invention provides a nucleic acid that is an allelic variant of *dptA, dptB, dptC, dptD* or *dptH*. In an even more preferred embodiment, the allelic variant is a variant of a gene, wherein the gene encodes DptA, DptB, DptC, DptD or DptH. In another preferred embodiment, the allelic variant is a variant of a gene that encodes a polypeptide comprising an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8. In a yet more preferred embodiment, the allelic variant is a variant of a gene, wherein the gene has the nucleic acid sequence of SEQ ID NOS: 10, 12, 14, 3 or 6. An allelic variant of *dptH* or the thioesterase of *dptD* preferably encodes a thioesterase with the same or similar enzymatic activity compared to that of the polypeptide having the amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or has the amino acid sequence of SEQ ID NO: 8. An allelic variant of *dptA*, *dptB*, *dptC* or *dptD* preferably encodes a polypeptide having the same activity as the daptomycin NRPS having the amino acid sequences of SEQ ID NOS: 9, 11, 13 or 7, respectively. In another embodiment, the invention provides an allelic variant of a nucleic acid molecule that encodes SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or to a nucleic acid molecule comprising the nucleic acid sequence SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. In a preferred embodiment, the allelic variant encodes a polypeptide having the same biological activity of the polypeptide; e.g., it encodes a polypeptide having ABC-transporter activity.

A further object of the invention is to provide a nucleic acid molecule that comprises a part of a nucleic acid sequence of the instant invention. The invention provides a part of a nucleic acid molecule encoding a daptomycin NRPS, a subunit thereof, a thioesterase from a daptomycin biosynthetic gene cluster, or a part of a nucleic acid molecule that comprises an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or, preferably, B12:03A05. The invention also provides a part of a selectively-hybridizing or homologous nucleic acid molecule, as described above. The invention provides a part of an allelic variant of a nucleic acid molecule, as described above. A part comprises at least 10 nucleotides, more preferably at least 15, 20, 25, 30, 35, 40, 50, 100, 150, 200, 250 or 300 nucleotides. The maximum size of a nucleic acid part is one nucleotide shorter than the entire nucleic acid molecule, if the nucleic acid molecule encodes more than one gene, or is one nucleotide shorter than the nucleic acid molecule encoding the full-length protein, if the nucleic acid molecule encodes a single polypeptide.

In another aspect, the hybridizing or homologous nucleic acid molecule, the allelic variant, or the part of the nucleic acid molecule encodes a polypeptide that has the same biological activity as the native (wild-type) polypeptide.

In another aspect, the invention provides a nucleic acid molecule that encodes a fusion protein, a homologous protein, a polypeptide fragment, a mutein or a polypeptide analog, as described below.

A nucleic acid molecule of this invention may encode a single polypeptide or multiple polypeptides. In one embodiment, the invention provides a nucleic acid molecule that encodes multiple, translationally coupled polypeptides, e.g., a nucleic acid molecule that encodes DptA, DptB, DptC and DptD. The invention also provides a nucleic acid molecule that encodes a single polypeptide derived from *S. roseosporus*, e.g., DptA, DptB, DptC or DptD, or a polypeptide fragment, mutein, fusion protein, polypeptide analog or homologous protein thereof. The invention also provides nucleic acid sequences, such as expression control sequences, that are not associated with other *S. roseosporus* sequences.

In one embodiment, the nucleic acid molecule may not consist of any one or more of the plasmids or cosmids designated pRHB152, pRHB153, pRHB154, pRHB155, pRHB157, pRHB159, pRHB160, pRHB161, pRHB162, pRHB166, pRHB168, pRHB169, pRHB170, pRHB172, pRHB173, pRHB174, pRHB599, pRHB602, pRHB603, pRHB613, pRHB614, pRHB680, pRHB678 or pRHB588 by McHenney et al., J. Bacteriol. 180: 143-151 (1998), herein incorporated by reference in its entirety. In another embodiment, the nucleic acid molecule may not consist of the nucleic acid sequence derived from *S. roseosporus* (the *S. roseosporus* insert) in any one of the above-mentioned plasmids or cosmids. In another embodiment, the nucleic acid molecule may not be the nucleic acid molecule may not consist of a vector into which the *S. roseosporus* insert from any one of the above-mentioned plasmids or cosmids has been inserted, wherein the vector comprises no other *S. roseosporus* sequences.

In another embodiment, the invention provides a nucleic acid molecule comprising one or more expression control sequences from a gene comprising a nucleic acid sequence that encodes a thioesterase or daptomycin NRPS from the daptomycin biosynthetic gene cluster. In a preferred embodiment, the nucleic acid molecule comprises a part or all of the expression control sequences of the daptomycin NRPS or *dptH*. In a yet more preferred embodiment, the nucleic acid molecule comprises all or a part of SEQ ID NO: 2 or SEQ ID NO: 5. In another preferred embodiment, the nucleic acid molecule comprises an expression control sequence from an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or, preferably, B12:03A05. Without wishing to be bound by any theory, it is thought that the nucleic acid sequence upstream of *dptA* in the daptomycin biosynthetic gene cluster (SEQ ID NO: 2) comprises the native expression control sequences for *dptA, dptB, dptC* and *dptD*. Further, it is thought that a single transcript for *dptA, dptB, dptC* and *dptD* is generated and that expression of DptA, DptB, DptC and DptD are translationally coupled.

In a preferred embodiment, the entire expression control sequence of a gene comprising a nucleic acid sequence that encodes a daptomycin NRPS and/or a thioesterase from the daptomycin biosynthetic gene cluster is used to control transcription. In another embodiment, only a part of the expression control sequence of a gene comprising a nucleic acid sequence that encodes a daptomycin NRPS and/or a thioesterase from the daptomycin biosynthetic gene cluster is used to control transcription. One having ordinary skill in the art may determine which part(s) of the gene to use to control transcription using methods known in the art. For instance, one may ligate a nucleic acid sequence comprising all or a part of an expression control sequence of a daptomycin NRPS and/or a thioesterase gene into a vector comprising a reporter gene. Examples of such reporter genes include, without limitation, chloramphenicol acetyltransferase (CAT), luciferase, green fluorescent protein, β-galactosidase and the like. The nucleic acid molecule comprising the expression control sequence is ligated into the vector such that it can act as a promoter or enhancer of the reporter gene. The vector is introduced into a host cell and expression is induced. Then, one may assay for the production of the reporter gene product to determine if the part(s) of the expression control sequence is sufficient to activate or regulate transcription. Methods of determining whether a nucleic acid sequence is sufficient to regulate transcription are routine and well-known in the art. See, e.g., Ausubel et al., *supra*.

A nucleic acid molecule comprising all or a part of an expression control sequence described herein, or multiple copies of these expression control sequences or parts thereof, may be operatively linked to a second nucleic acid molecule to regulate the transcription of the second nucleic acid molecule. In one embodiment, the invention provides a nucleic acid molecule comprising the expression control sequences operatively linked to a heterologous nucleic acid molecule, such as a nucleic acid molecule that encodes a polypeptide not usually expressed by *S. roseosporus*. In another preferred embodiment, the nucleic acid molecule comprising the expression control sequences is inserted into a vector, preferably a bacterial vector. In a more preferred embodiment, the vector is introduced into a bacterial host cell, more preferably into a *Streptomyces* or *E. coli*, and even more preferably into a *S. roseosporus*, *S. lividans* or *S. fradiae* host cell.

The invention also provides a nucleic acid sequence comprising the expression control sequence from *S. roseosporus* as described herein operatively linked to a nucleic acid sequence encoding a polypeptide involved in a daptomycin NRPS, a thioesterase derived from the daptomycin biosynthetic gene cluster, or a nucleic acid molecule from a BAC clone or part there as described herein. The expression control sequence may be operatively linked to a nucleic acid molecule encoding DptA, DptB, DptC, DptD or DptH, to a nucleic acid molecule encoding a polypeptide derived from the *S. roseosporus* sequences from a BAC clone of the invention, preferably B12:03A05, or to a nucleic acid molecule encoding a fragment, homologous protein, mutein, analog, derivative or fusion protein thereof. The expression control sequence may be operatively linked to a nucleic acid sequence encoding a polypeptide comprising an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8, or to a fragment thereof. Preferably, the expression control sequence is operatively linked to the coding region of one or more of *dptA, dptB, dptC, dptD* or *dptH*. In a more preferred embodiment, the expression control sequence is operatively linked to a nucleic acid sequence selected from SEQ ID NOS: 10, 12, 14, 3 or 6, or to a part thereof. The invention also provides an expression control sequence operatively linked to the coding region of a polypeptide comprising an amino acid sequence SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or to a nucleic acid molecule comprising the nucleic acid sequence SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102..

In another embodiment, the invention provides a nucleic acid molecule comprising one or more expression control sequences that directs the transcription of a nucleic acid molecule encoding a daptomycin NRPS, a subunit, module or domain thereof, a thioesterase, or a nucleic acid molecule encoding a polypeptide derived from the *S. roseosporus* sequences from a BAC clone of the invention, wherein the expression control sequence(s) are not derived from a daptomycin biosynthetic gene cluster. Examples of suitable expression control sequences are provided *infra*.

Expression Vectors, Host Cells and Recombinant Methods of Producing Polypeptides

Nucleic acid sequences may be expressed by operatively linking them to an expression control sequence in an appropriate expression vector and employing that expression vector to transform an appropriate unicellular host. Expression control sequences are sequences which control the transcription, post-transcriptional events and translation of nucleic acid sequences. Such operative linking of a nucleic sequence of this invention to an expression control sequence, of course, includes, if not already part of the nucleic acid sequence, the provision of a translation initiation codon, ATG or GTG, in the correct reading frame upstream of the nucleic acid sequence.

A wide variety of host/expression vector combinations may be employed in expressing the nucleic acid sequences of this invention. Useful expression vectors, for example, may consist of segments of chromosomal, non-chromosomal and synthetic nucleic acid sequences.

In a preferred embodiment, bacterial host cells are used to express the nucleic acid molecules of the instant invention. Useful expression vectors for bacterial hosts include bacterial plasmids, such as those from *E. coli* or *Streptomyces*, including pBluescript, pGEX-2T, pUC vectors, col E1, pCR1, pBR322, pMB9 and their derivatives, wider host range plasmids, such as RP4, phage DNAs, e.g., the numerous derivatives of phage lambda, e.g., NM989, λGT10 and λGT11, and other phages, e.g., M13 and filamentous single stranded phage DNA. A preferred vector is a bacterial artificial chromosome (BAC). A more preferred vector is pStreptoBAC, as described in Example 2.

In other embodiments, eukaryotic host cells, such as yeast or mammalian cells, may be used. Yeast vectors include Yeast Integrating plasmids (*e.g.*, YIp5) and Yeast Replicating plasmids (the YRp and YEp series plasmids), Yeast centromere plasmids (the YCp series plasmids), pGPD-2, 2μ plasmids and derivatives thereof, and improved shuttle vectors such as those described in Gietz and Sugino, Gene, 74, pp. 527-34 (1988) (YIplac, YEplac and YCplac). Expression in mammalian cells can be achieved using a variety of plasmids, including pSV2, pBC12BI, and p91023, as well as lytic virus vectors (*e.g.*, vaccinia virus, adeno virus, and baculovirus), episomal virus vectors (*e.g.*, bovine papillomavirus), and retroviral vectors (*e.g.*, murine retroviruses). Useful vectors for insect cells include baculoviral vectors and pVL 941.

In addition, any of a wide variety of expression control sequences may be used in these vectors to express the DNA sequences of this invention. Such useful expression control sequences include the expression control sequences associated with structural genes of the foregoing expression vectors. Expression control sequences that control transcription include, e.g., promoters, enhancers and transcription termination sites. Expression control sequences in eukaryotic cells that control post-transcriptional events include splice donor and acceptor sites and sequences that modify the half-life of the transcribed RNA, e.g., sequences that direct poly(A) addition or binding sites for RNA-binding proteins. Expression control sequences that control translation include ribosome binding sites, sequences which direct targeted expression of the polypeptide to or within particular cellular compartments, and sequences in the 5' and 3' untranslated regions that modify the rate or efficiency of translation.

Examples of useful expression control sequences include, for example, the early and late promoters of SV40 or adenovirus, the lac system, the trp system, the TAC or TRC system, the T3 and T7 promoters, the major operator and promoter regions of phage lambda, the control regions of fd coat protein, the promoter for 3-phosphoglycerate kinase or other glycolytic enzymes, the promoters of acid phosphatase, e.g., Pho5, the promoters of the yeast α-mating system, the GAL1 or GAL10 promoters, and other constitutive and inducible promoter sequences known to control the expression of genes of prokaryotic or eukaryotic cells or their viruses, and various combinations thereof. Other expression control sequences include those from the daptomycin biosynthetic gene cluster, such as those described *supra*.

Preferred nucleic acid vectors also include a selectable or amplifiable marker gene and means for amplifying the copy number of the gene of interest. Such marker genes are well-known in the art. Nucleic acid vectors may also comprise stabilizing sequences (e.g., ori- or ARS-like sequences and telomere-like sequences), or may alternatively be designed to favor directed or non-directed integration into the host cell genome. Preferred marker genes and stabilizing sequences are disclosed in pStreptoBAC, which is described in Example 2. In a preferred embodiment, nucleic acid sequences of this invention are inserted in frame into an expression vector that allows high level expression of an RNA which encodes a protein comprising the encoded nucleic acid sequence of interest. Nucleic acid cloning and sequencing methods are well known to those of skill in the art and are described in an assortment of laboratory manuals, including Sambrook et al., *supra*, 1989; and Ausubel et al. Product information from manufacturers of biological, chemical and immunological reagents also provide useful information. Example 2 provides preferred nucleic acid cloning and sequencing methods.

Of course, not all vectors and expression control sequences will function equally well to express the nucleic acid sequences of this invention. Neither will all hosts function equally well with the same expression system. However, one of skill in the art may make a selection among these vectors, expression control sequences and hosts without undue experimentation and without departing from the scope of this invention. For example, in selecting a vector, the host must be considered because the vector must be replicated in it. The vector's copy number, the ability to control that copy number, the ability to control integration, if any, and the expression of any other proteins encoded by the vector, such as antibiotic or other selection markers, should also be considered.

In selecting an expression control sequence, a variety of factors should also be considered. These include, for example, the relative strength of the sequence, its controllability, and its compatibility with the nucleic acid sequence of this invention, particularly with regard to potential secondary structures. Unicellular hosts should be selected by consideration of their compatibility with the chosen vector, the toxicity of the product coded for by the nucleic acid sequences of this invention, their secretion characteristics, their ability to fold the polypeptide correctly, their fermentation or culture requirements, and the ease of purification from them of the products coded for by the nucleic acid sequences of this invention.

The recombinant nucleic acid molecules and more particularly, the expression vectors of this invention may be used to express the polypeptides of this invention as recombinant polypeptides in a heterologous host cell. The polypeptides of this invention may be full-length or less than full-length polypeptide fragments recombinantly expressed from the nucleic acid sequences according to this invention. Such polypeptides include analogs, derivatives and muteins that may or may not have biological activity. In a preferred embodiment, the polypeptides are expressed in a heterologous bacterial host cell. In a more preferred embodiment, the polypeptides are expressed in a heterologous *Streptomyces* host cell, still more preferably a *S. lividans* or *S. fradiae* host cell. See, e.g., Example 7, *infra*.

Transformation and other methods of introducing nucleic acids into a host cell (e.g., conjugation, protoplast transformation or fusion, transfection, electroporation, liposome delivery, membrane fusion techniques, high velocity DNA-coated pellets, viral infection and protoplast fusion) can be accomplished by a variety of methods which are well known in the art (see, for instance, Ausubel, *supra*, and Sambrook et al., *supra*). Bacterial, yeast, plant or mammalian cells are transformed or transfected with an expression vector, such as a plasmid, a cosmid, or the like, wherein the expression vector comprises the nucleic acid of interest. Alternatively, the cells may be infected by a viral expression vector comprising the nucleic acid of interest. Depending upon the host cell, vector, and method of transformation used, transient or stable expression of the polypeptide will be constitutive or inducible. One having ordinary skill in the art will be able to decide whether to express a polypeptide transiently or in a stable manner, and whether to express the protein constitutively or inducibly.

A wide variety of unicellular host cells are useful in expressing the DNA sequences of this invention. These hosts may include well known eukaryotic and prokaryotic hosts, such as strains of *E. coli*, *Pseudomonas*, *Bacillus*, *Streptomyces*, fungi, yeast, insect cells such as *Spodoptera frugiperda* (SF9), animal cells such as CHO, BHK, MDCK and various murine cells, e.g., 3T3 and WEHI cells, African green monkey cells such as COS 1, COS 7, BSC 1, BSC 40, and BMT 10, and human cells such as VERO, WI38, and HeLa cells, as well as plant cells in tissue culture. In a preferred embodiment, the host cell is *Streptomyces*. In a more preferred embodiment, the host cell is *S. roseosporus*, *S. lividans* or *S. fradiae*.

Particular details of the transfection, expression and purification of recombinant proteins are well documented and are understood by those of skill in the art. Further details on the various technical aspects of each of the steps used in recombinant production of foreign genes in bacterial cell expression systems can be found in a number of texts and laboratory manuals in the art. See, e.g., Ausubel et al., *supra*, and Sambrook et al., *supra*, and Kieser et al., *supra*, herein incorporated by reference.

Polypeptides

Thioesterases and Fragments Thereof

Another object of the invention is to provide a polypeptide derived from a thioesterase involved in daptomycin synthesis. In one embodiment, the polypeptide is derived from a daptomycin biosynthetic gene cluster. In a preferred embodiment, the polypeptide is derived from an integral or free thioesterase. In a more preferred embodiment, the polypeptide comprises the thioesterase domain of DptD or the amino acid sequence of DptH. In an even more preferred embodiment, the polypeptide comprises the amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or the amino acid sequence of SEQ ID NO: 8. The polypeptide derived from a thioesterase may also be encoded by an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably from B12:03A05. A polypeptide as defined herein may be produced recombinantly, as discussed *supra*, may be isolated from a cell that naturally expresses the protein, or may be chemically synthesized following the teachings of the specification and using methods well known to those having ordinary skill in the art. See, e.g., Examples 3-6.

The polypeptide may comprise a fragment of a thioesterase as defined herein. A polypeptide that comprises only a part or fragment of the entire thioesterase may or may not encode a polypeptide that has thioesterase activity. A polypeptide that does not have thioesterase activity, whether it is a fragment, analog, mutein, homologous protein or derivative, is nevertheless useful, especially for immunizing animals to prepare anti-thioesterase antibodies. However, in a preferred embodiment, the part or fragment encodes a polypeptide having thioesterase activity. Methods of determining whether a polypeptide has thioesterase activity are described *infra*. Further, in a preferred embodiment, the fragment comprises an amino acid sequence comprising the GXSXG thioesterase motif (see Example 3). In a more preferred embodiment, the fragment comprises an amino acid sequence comprising the thioesterase motif GWSFG or GTSLG, which are derived from the thioesterase domain of SEQ ID NO: 7 or the amino acid sequence of SEQ ID NO: 8, respectively.

One can produce fragments of a polypeptide encoding a thioesterase by truncating the DNA encoding the thioesterase and then expressing it recombinantly. Alternatively, one can produce a fragment by chemically synthesizing a portion of the full-length polypeptide. One may also produce a fragment by enzymatically cleaving either a recombinant polypeptide or an isolated naturally-occurring polypeptide. Methods of producing polypeptide fragments are well-known in the art (see, e.g., Sambrook et al. and Ausubel et al., *supra*). In one embodiment, a polypeptide comprising only a part or fragment of a thioesterase may be produced by chemical or enzymatic cleavage of a thioesterase. In a preferred embodiment, a polypeptide fragment is produced by expressing a nucleic acid molecule encoding a fragment of the thioesterase in a host cell.

Daptomycin NRPS Polypeptides, and Subunits and Fragments Thereof

Another object of the invention is to provide a polypeptide derived from a daptomycin NRPS or subunit thereof. The daptomycin NRPS comprises the subunits DptA, DptB, DptC and DptD. As discussed in greater detail in Examples 3-6 below, each subunit comprises a number of modules that bind and activate specific building block substrates and to catalyze peptide chain formation and elongation. Further, each module comprises a number of domains that participate in condensation, adenylation and thiolation. In addition, some modules comprise a epimerization domain, discussed in greater detail in Example 6. DptD also comprises a thioesterase domain, as discussed *supra* and in Example 5.

In one embodiment, the polypeptide an amino acid sequence from DptA, DptB, DptC and/or DptD. In an even more preferred embodiment, the polypeptide comprises an amino acid sequence SEQ ID NOS: 9, 11, 13 or 7. A daptomycin NRPS polypeptide may also be encoded by an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably from B12:03A05. A polypeptide as defined herein may be produced recombinantly, as discussed *supra*, may be isolated from a cell that naturally expresses the protein, or may be chemically synthesized following the teachings of the specification and using methods well known to those having ordinary skill in the art. See, e.g., Examples 3-6 regarding amino acid sequences as well as modules and domains of DptA, DptB, DptC and DptD.

The polypeptide may comprise a fragment of a daptomycin NRPS as defined herein. In one embodiment, a fragment comprises one or more complete modules of a daptomycin NRPS subunit. In another embodiment, a fragment comprises one or more domains of a daptomycin NRPS subunit. In yet another embodiment, a fragment may not comprise a complete domain or module but may comprise only a part of one or more domains or modules. A polypeptide that does not comprise a full domain or module of a daptomycin NRPS, whether it is a fragment, analog, mutein, homologous protein or derivative, is nevertheless useful, especially for immunizing animals to prepare anti-thioesterase antibodies. In a more preferred embodiment, the fragment comprises an amino acid sequence comprising at least that part of an adenylation domain that is required for binding to an amino acid. This part of the domain is delimited by the amino acid pocket code of a particular adenylation domain, as discussed below in Example 5.

As discussed above, one can produce fragments of a polypeptide of the invention recombinantly, by chemical synthesis or by enzymatic cleavage.

Polypeptides from S. roseosporus BAC Clones

Another object of the invention is to provide a polypeptide encoded by a nucleic acid molecule or part thereof from a *S. roseosporus* BAC clone of the invention. In one embodiment, the invention provides a polypeptide encoded by a nucleic acid molecule or part thereof from 1G05, 06A12, 12F06, 18H04, 20C09 or, preferably, B12:03A05. In a preferred embodiment, the invention provides a polypeptide comprising an amino acid sequence SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or encoded by a nucleic acid molecule comprising the nucleic acid sequence SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. In another preferred embodiment, the invention provides a polypeptide that is DptE or DptF, a polypeptide having an amino acid sequence of SEQ ID NO: 15 or SEQ ID NO: 17, or encoded by *dptE* or *dptF*, or encoded by a nucleic acid sequence of SEQ ID NO: 16 or SEQ ID NO: 18. In another preferred embodiment, the invention provides an ABC transporter comprising an amino acid sequence SEQ ID NOS: 19, 21, 29, 45, 47, 49, 63, 67, 75 and 77, or encoded by a nucleic acid sequence of SEQ ID NOS: 20, 22, 30, 46, 48, 50, 64, 68, 76 or 78. In another preferred embodiment, the invention provides a polypeptide that is an oxidoreductase, such as a dehydrogenase; a transcriptional regulator involved in antibiotic resistance; NovABC-related polypeptides, which are involved in the biosynthesis of novobiocin, an antimicrobial agent; a monooxygenase; an acyl CoA thioesterase; a DNA helicase; or a DNA ligase, such as provided by a polypeptide having an amino acid sequence selected from SEQ ID NOS: 23, 25, 27, 29, 33, 35, 37, 91, 93, 97 and 99. In another preferred embodiment, the invention provides a polypeptide that is highly homologous to a *Streptomyces* polypeptide, such as provided by a polypeptide having an amino acid sequence selected from SEQ ID NOS: 61, 65, 69, 79, 81, 83, 85, 87, 95 and 101. A polypeptide as defined herein may be produced recombinantly, as discussed *supra*, may be isolated from a cell that naturally expresses the protein, or may be chemically synthesized following the teachings of the specification and using methods well known to those having ordinary skill in the art. See, e.g., Example X. The invention also provides a polypeptide that comprises a fragment of a nucleic acid molecule that encodes a polypeptide from a BAC clone, as defined herein. As discussed above, one can produce fragments of a polypeptide of the invention recombinantly, by chemical synthesis or by enzymatic cleavage.

Muteins, Homologous Proteins, Allelic Variants, Analogs and Derivatives

Another object of the invention is to provide polypeptides that are mutant proteins (muteins), fusion proteins, homologous proteins or allelic variants of the daptomycin NRPS, subunits thereof, thioesterases or the polypeptides encoded by the *S. roseosporus* BAC nucleic acid molecules or parts thereof provided herein. A mutant thioesterase may have the same or different enzymatic activity compared to a naturally-occurring thioesterase and comprises at least one amino acid insertion, duplication, deletion, rearrangement or substitution compared to the amino acid sequence of a native protein. In one embodiment, the mutein has the same or a decreased thioesterase activity compared to a naturally-occurring thioesterase. In another embodiment, the mutant thioesterase has an increased thioesterase activity compared to a naturally-occurring thioesterase. In a preferred embodiment, muteins of thioesterases of a daptomycin biosynthetic gene cluster may be used to alter thioesterase activity. See, e.g., Examples 12 and 16. In another embodiment, a mutant daptomycin NRPS or subunit thereof may have the same or different amino acid specificity, thiolation activity, condensation activity, or, if present, epimerization activity, as a naturally-occurring daptomycin NRPS. Daptomycin NRPS muteins may be used to alter amino acid recognition, binding, epimerization or other catalytic properties of an NRPS. See, e.g., Examples 12 and 16. Similarly, a mutein of a polypeptide encoded by the *S. roseosporus* BAC nucleic acid molecule of the invention may have a similar biological activity or a different one, but preferably has a similar biological activity.

A mutein of the invention may be produced by isolation from a naturally-occurring mutant microorganism or from a microorganism that has been experimentally mutagenized, may be produced by chemical manipulation of a polypeptide, or may be produced from a host cell comprising an altered nucleic acid molecule. In a preferred embodiment, the mutein is produced from a host cell comprising an altered nucleic acid molecule. Muteins may also be produced chemically by altering the amino acid residue to another amino acid residue using synthetic or semi-synthetic chemical techniques. One may produce muteins of a polypeptide by introducing mutations into the nucleic acid sequence encoding a daptomycin NRPS, subunit thereof or a thioesterase, or into a *S. roseosporus* BAC nucleic acid molecule, and then expressing it recombinantly. These mutations may be targeted, in which particular encoded amino acids are altered, or may be untargeted, in which random encoded amino acids within the polypeptide are altered. Muteins with random amino acid alterations can be screened for a particular biological activity, such as thioesterase activity, amino acid specificity, thiolation activity, epimerization activity, or condensation activity, as described below. Muteins may also be screened, e.g., for oxidoreductase activity, ABC transporter activity, monooxygenase activity, or DNA ligase or helicase activity using methods known in the art. Multiple random mutations can be introduced into the gene by methods well-known to the art, e.g., by error-prone PCR, shuffling, oligonucleotide-directed mutagenesis, assembly PCR, sexual PCR mutagenesis, *in vivo* mutagenesis, cassette mutagenesis, recursive ensemble mutagenesis, exponential ensemble mutagenesis and site-specific mutagenesis. Methods of producing muteins with targeted or random amino acid alterations are well known in the art. See, e.g., Sambrook et al., *supra*, Ausubel et al., *supra,* U.S. Pat. No. 5,223,408, and the references discussed *supra*, each herein incorporated by reference.

The invention also provides a polypeptide that is homologous to a daptomycin NRPS, subunit thereof, a thioesterase from a daptomycin biosynthetic gene cluster, or to a polypeptide encoded by a *S. roseosporus* BAC nucleic acid molecule as described herein. In one embodiment, the polypeptide is homologous to the thioesterase domain of DptD or to DptH, or to a polypeptide encoded by the thioesterase domain of *dptD* or by *dptH*. In a preferred embodiment, the polypeptide is homologous to a thioesterase having the amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or having the amino acid sequence of SEQ ID NO: 8. In another embodiment, the polypeptide is homologous to DptA, DptB, DptC or DptD, or to a polypeptide encoded by *dptA, dptB, dptC* or *dptD*. In a more preferred embodiment, the polypeptide is homologous to a polypeptide having the amino acid sequence of SEQ ID NO: 9, 11, 13 or 3. The invention also provides a polypeptide that is homologous to a polypeptide encoded by a nucleic acid molecule from a *S. roseosporus* BAC clone described herein, e.g., 1G05, 06A12, 12F06, 18H04, 20C09 or, preferably, B12:03A05. In a preferred embodiment, the invention provides a polypeptide homologous to a polypeptide comprising an amino acid sequence of SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or encoded by a nucleic acid molecule comprising a nucleic acid sequence selected from SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102.

In a preferred embodiment, the homologous polypeptide is one that exhibits significant sequence identity to a polypeptide of the invention. In a more preferred embodiment, the homologous polypeptide is one that exhibits at least 50%, 60%, 70%, or 80% sequence identity to a polypeptide comprising an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8 or SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101. In an even more preferred embodiment, the homologous polypeptide is one that exhibits at least 85%,90%, 95%, 96%, 97%, 98% or 99% sequence identity to a polypeptide comprising an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8 or SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101.

The homologous protein may be a naturally-occurring one that is derived from another species, especially one derived from another *Streptomyces* species, or one derived from another *Streptomyces roseosporus* strain, wherein the homologous protein comprises an amino acid sequence that exhibits significant sequence identity to that of SEQ ID NOS: 9, 11, 13, 7 or 8 or SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101. The naturally-occurring homologous protein may be isolated directly from the other species or strain. Alternatively, the nucleic acid molecule encoding the naturally-occurring homologous protein may be isolated and used to express the homologous protein recombinantly. In another embodiment, the homologous protein may be one that is experimentally produced by random mutation of a nucleic acid molecule and subsequent expression of the nucleic acid molecule. In another embodiment, the homologous protein may be one that is experimentally produced by directed mutation of one or more codons to alter the encoded amino acid of the polypeptide.

In another embodiment, the invention provides a polypeptide encoded by an allelic variant of a gene encoding a thioesterase from a daptomycin biosynthetic gene cluster, or a daptomycin NRPS or subunit thereof. In a preferred embodiment, the invention provides a polypeptide encoded by an allelic variant of *dptA*, *dptB*, *dptC*, *dptD* or *dptH*. In an even more preferred embodiment, the polypeptide is encoded by an allelic variant of a gene that encodes a polypeptide having the amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8. In a yet more preferred embodiment, the polypeptide is encoded by an allelic variant of a gene, wherein the gene has the nucleic acid sequence of SEQ ID NOS: 10, 12, 14, 3 or 6. An allelic variant may have the same or different biological activity as the thioesterase, daptomycin NRPS or subunit thereof, described herein. In a preferred embodiment, an allelic variant is derived from another species of *Streptomyces*, even more preferably from a strain of *Streptomyces roseosporus*. In another embodiment, the invention provides a polypeptide encoded by an allelic variant of an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably from B12:03A05. In a preferred embodiment, the polypeptide is encoded by an allelic variant of a gene that encodes a polypeptide having the amino acid sequence of SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101, or that is encoded by an allelic variant of a gene, wherein the gene has a nucleic acid sequence of SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102.

In another embodiment, the invention provides a derivative of a polypeptide of the invention. In a preferred embodiment, the derivative has been acetylated, carboxylated, phosphorylated, glycosylated or ubiquitinated. In another preferred embodiment, the derivative has been labeled with, e.g., radioactive isotopes such as ^{125}I, ^{32}P, ^{35}S, and ^{3}H. In another preferred embodiment, the derivative has been labeled with fluorophores, chemiluminescent agents, enzymes, and antiligands that can serve as specific binding pair members for a labeled ligand. In a preferred embodiment, the polypeptide is a thioesterase involved in the biosynthesis of daptomycin. In an even more preferred embodiment, the polypeptide comprises the thioesterase domain of DptD or comprises the amino acid sequence of DptH, or is a thioesterase encoded by the thioesterase-encoding domain of *dptD* or by *dptH*. In another preferred embodiment, the polypeptide is a daptomycin NRPS or subunit thereof, more preferably DptA, DptB, DptC or DptD, even more preferably a polypeptide encoded by *dptA, dptB, dptC* or *dptD*. In a yet more preferred embodiment, the polypeptide has an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8 or is a mutein, allelic variant, homologous protein or fragment thereof. Preferably, a thioesterase derivative has a thioesterase activity that is the same or similar to a thioesterase involved in the biosynthesis of daptomycin, more preferably, the derivative has a thioesterase activity that is the same or similar to a thioesterase having an amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or having the amino acid sequence of SEQ ID NO: 8. In another preferred embodiment, a daptomycin NRPS or NRPS subunit derivative has the same or similar activity as a naturally-occurring daptomycin NRPS or subunit thereof. In yet another embodiment, the derivative is derived from a polypeptide encoded by a nucleic acid molecule from a *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or, preferably, B12:03A05. In a preferred embodiment, the derivative is derived from a polypeptide having an amino acid sequence of SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101, or that is encoded by a gene having a nucleic acid sequence of SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102.

The invention also provides non-peptide analogs. In a preferred embodiment, the non-peptide analog is structurally similar to a thioesterase involved in daptomycin synthesis, to a daptomycin NRPS or subunit thereof, or to a polypeptide encoded by a nucleic acid molecule from an *S. roseosporus* BAC clone, but in which one or more peptide linkages is replaced by a linkage selected from the group consisting of --CH_2NH--, --CH_2S--, --CH_2-CH_2--, --CH=CH--(cis and trans), --$COCH_2$--, --CH(OH)CH_2-- and –CH_2SO--. In another embodiment, the non-peptide analog comprises substitution of one or more amino acids of a thioesterase or daptomycin NRPS or subunit thereof with a D-amino acid of the same type in order to generate more stable peptides. Preferably, both a non-peptide and a peptide analog has a biological activity that is the same or similar to the naturally-occurring polypeptide involved in the biosynthesis of daptomycin, more preferably, the analog has a biological activity that is the same or similar to the polypeptide having an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8. The invention also provides analogs of polypeptides encoded by an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably from B12:03A05. The invention provides an analog of a polypeptide having an amino acid sequence of SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101, or that is encoded by a gene having a nucleic acid sequence of SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102.

Fusion Proteins

The polypeptides of this invention may be fused to other molecules, such as genetic, enzymatic or chemical or immunological markers such as epitope tags. Fusion partners include, *inter alia*, *myc*, hemagglutinin (HA), GST, immunoglobulins, β-galactosidase, biotin trpE, protein A, β-lactamase, α-amylase, maltose binding protein, alcohol dehydrogenase, polyhistidine (for example, six histidine at the amino and/or carboxyl terminus of the polypeptide), lacZ, green fluorescent protein (GFP), yeast α mating factor, GAL4 transcription activation or DNA binding domain, luciferase, and serum proteins such as ovalbumin, albumin and the constant domain of IgG. See, e.g., Godowski et al., 1988, and Ausubel et al., *supra*. Fusion proteins may also contain sites for specific enzymatic cleavage, such as a site that is recognized by enzymes such as Factor XIII, trypsin, pepsin, or any other enzyme known in the art. Fusion proteins will typically be made by either recombinant nucleic acid methods, as described above, chemically synthesized using techniques such as those described in Merrifield, 1963, herein incorporated by reference, or produced by chemical cross-linking.

Tagged fusion proteins permit easy localization, screening and specific binding via the epitope or enzyme tag. See Ausubel, 1991, Chapter 16. Some tags allow the protein of interest to be displayed on the surface of a phagemid, such as M13, which is useful for panning agents that may bind to the desired protein targets. Another advantage of fusion proteins is that an epitope or enzyme tag can simplify purification. These fusion proteins may be purified, often in a single step, by affinity chromatography. For example, a His^6 tagged protein can be purified on a Ni affinity column and a GST fusion protein can be purified on a glutathione affinity column. Similarly, a fusion protein comprising the Fc domain of IgG can be purified on a Protein A or Protein G column and a fusion protein comprising an epitope tag such as myc can be purified using an immunoaffinity column containing an anti-c-myc antibody. It is preferable that the epitope tag be separated from the protein encoded by the nucleic acid molecule of the invention by an enzymatic cleavage site that can be cleaved after purification.

A second advantage of fusion proteins is that the epitope tag can be used to bind the fusion protein to a plate or column through an affinity linkage for screening targets.

Therefore, in another aspect, the invention provides a fusion protein comprising all or a part of a thioesterase derived from a daptomycin biosynthetic gene cluster and provides a nucleic acid molecule that encodes such a fusion protein. Another aspect provides a fusion protein comprising all or a part of a daptomycin NRPS or subunit thereof and provides a nucleic acid molecule encoding such a protein. See, e.g., Examples 11-16. The invention also provides a fusion protein comprising all or part of a polypeptide encoded by a nucleic acid molecule from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05. In a preferred embodiment, the fusion protein comprises all or a part of a polypeptide encoded by one or more of *dptA, dptB, dptC, dptD* or *dptH*. In another preferred embodiment, the fusion protein comprises a polypeptide encoded by a nucleic acid molecule that selectively hybridizes to *dptA, dptB, dptC, dptD* or *dptH*. In a more preferred embodiment, the fusion protein comprises a polypeptide having an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8, or comprises a polypeptide that is a fragment, mutein, homologous protein, derivative or analog thereof. In an even more preferred embodiment, the nucleic acid molecule encoding the fusion protein comprises all or part of the nucleic acid sequence of SEQ ID NOS: 10, 12, 14, 3 or 6, or comprises all or part of a nucleic acid sequence that selectively hybridizes or is homologous to a nucleic acid molecule comprising said nucleic acid sequence. The invention also provides fusion proteins comprising polypeptide sequences encoded by an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably from B12:03A05. The invention provides a fusion protein comprising a polypeptide having an amino acid sequence of SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101, or comprising a polypeptide that is a fragment, mutein, homologous protein, derivative or analog thereof. The invention also provides a fusion protein comprising a polypeptide encoded by SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102, or comprising all or part of a nucleic acid sequence that selectively hybridizes or is homologous to a nucleic acid molecule comprising said nucleic acid sequence.

In one aspect of the invention, the fusion protein that comprises all or a part of a thioesterase derived from a daptomycin biosynthetic gene cluster comprises other modules (including heterologous or hybrid modules) from a polypeptide involved in non-ribosomal protein synthesis. See, e.g., Examples 12E, G and H and Example 16. In another preferred embodiment, the fusion protein comprises one or more amino acid sequences that encode thioesterases, wherein the thioesterases may be identical to one another or may be different. See, e.g., Examples 11E-G (duplication of daptomycin thioesterase genes), Example 12 (producing modified NRPS thioesterase fusion proteins) and Example 16 (producing free thioesterase fusion proteins).

In another embodiment, the invention provides a fusion protein that is a hybrid of amino acid sequences from two or more different thioesterases and a nucleic acid molecule that encodes such a fusion protein. The hybrid fusion protein may consist of two, three or more portions of different thioesterases. The hybrid thioesterase may have a different or the same specificity.

Methods to Assay Thioesterase and Daptomycin NRPS Activity

There are a number of methods known in the art to determine whether a fragment, mutein, homologous protein, analog, derivative or fusion protein of a thioesterase has the same, enhanced or decreased biological activity as a wild-type thioesterase polypeptide. In one embodiment, a thioesterase assay which monitors cleavage of a suitable thioester bond and/or release of a corresponding product is performed *in vitro*. Any of a number of thioesterase assays well-known in the art may be used, including those which use photo- or radio-labeled substrates.

In a preferred embodiment, thioesterase activity associated with peptide synthesis by a NRPS is determined using cellular assays. For example, a nucleic acid molecule encoding a fragment, mutein, homologous protein or fusion protein may be introduced into a bacterial cell comprising a daptomycin biosynthetic gene cluster absent one or both of the thioesterase domains of *dptD* or *dptH*. Alternatively, the nucleic acid molecule may be introduced into a bacterial cell comprising a different biosynthetic gene cluster that produces a different compound, e.g., a different lipopeptide. In a preferred embodiment, the bacterial cell may be *S. lividans*. The nucleic acid molecule may be introduced into the bacterial cell by any method known in the art, including conjugation, transformation, electroporation, protoplast fusion or the like. The bacterial cell comprising the nucleic acid molecule is incubated under conditions in which the polypeptide encoded by the nucleic acid molecule is expressed. After incubation, the bacterial cells may be analyzed by, e.g., HPLC and/or LC/MS, to determine if the bacterial cells produce the desired lipopeptide. See, e.g., the method of expressing daptomycin described in Examples 7- 9, *infra*. When the thioesterase activity is associated with synthesis of a peptide having an anti-cell growth property (e.g., an antibiotic, antifungal, antiviral or antimitotic agent) an assay such as that described in Example 15 may be used. See Example 17.

Alternatively, a fragment, mutein, homologous protein, analog, derivative or fusion protein of a thioesterase may be introduced into a cell, particularly a bacterial cell, comprising a daptomycin biosynthetic gene cluster absent one or both of the thioesterase domain of *dptD* or *dptH*. After incubation, the bacterial cells may be analyzed by, e.g., HPLC and/or LC/MS, as described in Example 7, to determine if the bacterial cells produce the desired lipopeptide. The same method can be used with a cell comprising a different biosynthetic gene cluster that produces a different compound, e.g., a different lipopeptide.

In a preferred embodiment, a fragment, mutein, homologous protein, analog, derivative or fusion protein comprises an amino acid sequence comprising the GXSXG thioesterase motif (see Example 3). In a more preferred embodiment, a fragment, mutein, homologous protein, analog or derivative comprises an amino acid sequence comprising the thioesterase motif GWSFG or GTSLG, which are derived from SEQ ID NO: 7 and SEQ ID NO: 8, respectively.

Similar methods known in the art may be used to determine whether a fragment, mutein, homologous protein, analog, derivative or fusion protein of a daptomycin NRPS or subunit thereof has the same or different biological activity as a wild-type NRPS or subunit thereof.

Antibodies

The polypeptides encoded by the genes of this invention may be used to elicit polyclonal or monoclonal antibodies that bind to a polypeptide of this invention, as well as a fragment, mutein, homologous protein, analog, derivative or fusion protein thereof, using a variety of techniques well known to those of skill in the art. Antibodies directed against the polypeptides of this invention are immunoglobulin molecules or portions thereof that are immunologically reactive with the polypeptide of the present invention.

Antibodies directed against a polypeptide of the invention may be generated by immunization of a mammalian host. Such antibodies may be polyclonal or monoclonal. Preferably they are monoclonal. Methods to produce polyclonal and monoclonal antibodies are well known to those of skill in the art. For a review of such methods, see Harlow and Lane, Antibodies: A Laboratory Manual (1988) and Ausubel et al. *supra*, herein incorporated by reference. Determination of immunoreactivity with a polypeptide of the invention may be made by any of several methods well known in the art, including by immunoblot assay and ELISA.

Monoclonal antibodies with affinities of 10^{-8} M^{-1} or preferably 10^{-9} to 10^{-10} M^{-1} or stronger are typically made by standard procedures as described, e.g., in Harlow and Lane, 1988. Briefly, appropriate animals are selected and the desired immunization protocol followed. After the appropriate period of time, the spleens of such animals are excised and individual spleen cells fused, typically, to immortalized myeloma cells under appropriate selection conditions. Thereafter, the cells are clonally separated and the supernatants of each clone tested for their production of an appropriate antibody specific for the desired region of the antigen.

Other suitable techniques involve *in vitro* exposure of lymphocytes to the antigenic polypeptides, or alternatively, to selection of libraries of antibodies in phage or similar vectors. See Huse et al., 1989. The polypeptides and antibodies of the present invention may be used with or without modification. Frequently, polypeptides and antibodies will be labeled by joining, either covalently or non-covalently, a substance which provides for a detectable signal. A wide variety of labels and conjugation techniques are known and are reported extensively in both the scientific and patent literature. Suitable labels include radionuclides, enzymes, substrates, cofactors, inhibitors, fluorescent agents, chemiluminescent agents, magnetic particles and the like. Patents teaching the use of such labels include U.S. Patents 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149 and 4,366,241, herein incorporated by reference. Also, recombinant immunoglobulins may be produced (see U.S. Patent 4,816,567, herein incorporated by reference).

An antibody of this invention may also be a hybrid molecule formed from immunoglobulin sequences from different species (e.g., mouse and human) or from portions of immunoglobulin light and heavy chain sequences from the same species. An antibody may be a single-chain antibody or a humanized antibody. It may be a molecule that has multiple binding specificities, such as a bifunctional antibody prepared by any one of a number of techniques known to those of skill in the art including the production of hybrid hybridomas, disulfide exchange, chemical cross-linking, addition of peptide linkers between two monoclonal antibodies, the introduction of two sets of immunoglobulin heavy and light chains into a particular cell line, and so forth.

The antibodies of this invention may also be human monoclonal antibodies, for example those produced by immortalized human cells, by SCID-hu mice or other non-human animals capable of producing "human" antibodies, or by the expression of cloned human immunoglobulin genes. The preparation of humanized antibodies is taught by U.S. Pat. Nos. 5,777,085 and 5,789,554, herein incorporated by reference.

In sum, one of skill in the art, provided with the teachings of this invention, has available a variety of methods which may be used to alter the biological properties of the antibodies of this invention including methods which would increase or decrease the stability or half-life, immunogenicity, toxicity, affinity or yield of a given antibody molecule, or to alter it in any other way that may render it more suitable for a particular application.

In a preferred embodiment, an antibody of the present invention binds to a thioesterase involved in daptomycin synthesis or to a daptomycin NRPS or subunit thereof. In a more preferred embodiment, the antibody binds to a polypeptide encoded by *dptA*, *dptB*, *dptC*, *dptD* or *dptH*, or to a fragment thereof. In another preferred embodiment, the antibody binds to a polypeptide encoded by a nucleic acid molecule that selectively hybridizes to *dptA*, *dptB*, *dptC*, *dptD* or *dptH*. In a more preferred embodiment, the antibody binds to a polypeptide having an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8, or binds to a polypeptide that is fragment, mutein, homologous protein, derivative, analog or fusion protein thereof. In an even more preferred embodiment, the antibody binds to a polypeptide encoded by a nucleic acid molecule comprising all or part of the nucleic acid sequence of SEQ ID NOS: 10, 12, 14, 3 or 6. In another embodiment, the antibody binds to a polypeptide encoded by a nucleic acid molecule that comprises all or part of a nucleic acid sequence that selectively hybridizes or is homologous to a nucleic acid molecule comprising a nucleic acid sequence of SEQ ID NOS: 10, 12, 14, 3 or 6.

The invention provides an antibody that selectively binds to a polypeptide encoded by an *S. roseosporus* nucleic acid sequence from any one of BAC clones 01G05, 06A12, 12F06, 18H04, 20C09 or B12:03A05, preferably from B12:03A05. The polypeptide may comprise an amino acid sequence selected from SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101 or is encoded by a nucleic acid sequence SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. Preferably, the antibody selectively binds to a polypeptide comprising an amino acid sequence selected from SEQ ID NOS: 23, 25, 27, 29, 33, 35, 37, 91, 93, 97 and 99 or from SEQ ID NOS: 61, 65, 69, 79, 81, 83, 85, 87, 95 and 101. The invention also provides an antibody that selectively binds to a fragment, mutein, homologous protein, derivative, analog or fusion protein thereof.

Computer Readable Means

A further aspect of the invention is a computer readable means for storing the nucleic acid and amino acid sequences of the instant invention. In a preferred embodiment, the invention provides a computer readable means for storing all of the nucleic acid and amino acid sequences described herein, as the complete set of sequences or in any combination. The records of the computer readable means can be accessed for reading and display and for interface with a computer system for the application of programs allowing for the location of data upon a query for data meeting certain criteria, the comparison of sequences, the alignment or ordering of sequences meeting a set of criteria, and the like.

The nucleic acid and amino acid sequences of the invention are particularly useful as components in databases useful for search analyses as well as in sequence analysis algorithms. As used herein, the terms "nucleic acid sequences of the invention" and "amino acid sequences of the invention" mean any detectable chemical or physical characteristic of a polynucleotide or polypeptide of the invention that is or may be reduced to or stored in a computer readable form. These include, without limitation, chromatographic scan data or peak data, photographic data or scan data therefrom, and mass spectrographic data.

This invention provides computer readable media having stored thereon sequences of the invention. A computer readable medium may comprise one or more of the following: a nucleic acid sequence comprising a sequence of a nucleic acid sequence of the invention; an amino acid sequence comprising an amino acid sequence of the invention; a set of nucleic acid sequences wherein at least one of said sequences comprises the sequence of a nucleic acid sequence of the invention; a set of amino acid sequences wherein at least one of said sequences comprises the sequence of an amino acid sequence of the invention; a data set representing a nucleic acid sequence comprising the sequence of one or more nucleic acid sequences of the invention; a data set representing a nucleic acid sequence encoding an amino acid sequence comprising the sequence of an amino acid sequence of the invention; a set of nucleic acid sequences wherein at least one of said sequences comprises the sequence of a nucleic acid sequence of the invention; a set of amino acid sequences wherein at least one of said sequences comprises the sequence of an amino acid sequence of the invention; a data set representing a nucleic acid sequence comprising the sequence of a nucleic acid sequence of the invention; a data set representing a nucleic acid sequence encoding an amino acid sequence comprising the sequence of an amino acid sequence of the invention. The computer readable medium can be any composition of matter used to store information or data, including, for example, commercially available floppy disks, tapes, hard drives, compact disks, and video disks.

Also provided by the invention are methods for the analysis of character sequences, particularly genetic sequences. Preferred methods of sequence analysis include, for example, methods of sequence homology analysis, such as identity and similarity analysis, RNA structure analysis, sequence assembly, cladistic analysis, sequence motif analysis, open reading frame determination, nucleic acid base calling, and sequencing chromatogram peak analysis.

A computer-based method is provided for performing nucleic acid homology identification. This method comprises the steps of providing a nucleic acid sequence comprising the sequence a nucleic acid of the invention in a computer readable medium; and comparing said nucleic acid sequence to at least one nucleic acid or amino acid sequence to identify homology.

A computer-based method is also provided for performing amino acid homology identification, said method comprising the steps of: providing an amino acid sequence comprising the sequence of an amino acid of the invention in a computer readable medium; and comparing said an amino acid sequence to at least one nucleic acid or an amino acid sequence to identify homology.

A computer based method is still further provided for assembly of overlapping nucleic acid sequences into a single nucleic acid sequence, said method comprising the steps of: providing a first nucleic acid sequence comprising the sequence of a nucleic acid of the invention in a computer readable medium; and screening for at least one overlapping region between said first nucleic acid sequence and a second nucleic acid sequence.

Methods of Using Nucleic Acid Molecules as Probes and Primers

In one embodiment, a nucleic acid molecule of the invention may be used as a probe or primer to identify or amplify a nucleic acid molecule that selectively hybridizes to the nucleic acid molecule. In a preferred embodiment, the probe or primer is derived from a nucleic acid molecule encoding a daptomycin NRPS, subunit thereof or thioesterase from a daptomycin biosynthetic gene cluster. The probe or primer may also be derived from an expression control sequence derived from a daptomycin NRPS or thioesterase gene of a daptomycin biosynthetic gene cluster. In a preferred embodiment, the probe or primer is derived from *dptA, dptB, dptC, dptD* or *dptH*. In a more preferred embodiment, the probe or primer is derived from a nucleic acid molecule that encodes a polypeptide having an amino acid sequence of SEQ ID NOS: 9, 11, 13, 7 or 8. In a yet more preferred embodiment, the probe or primer is derived from a nucleic acid molecule that has a nucleic acid sequence of SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 or 102. In another embodiment, the probe or primer is derived from a nucleic acid sequence that encodes SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 or 101.

In general, a probe or primer is at least 10 nucleotides in length, more preferably at least 12, more preferably at least 14 and even more preferably at least 16 nucleotides in length. In an even more preferred embodiment, the probe or primer is at least 18 nucleotides in length, even more preferably at least 20 nucleotides and even more preferably at least 22 nucleotides in length. Primers and probes may also be longer in length. For instance, a probe or primer may be 25 nucleotides in length, or may be 30, 40 or 50 nucleotides in length. Methods of performing nucleic acid hybridization using oligonucleotide probes are well-known in the art. See, e.g., Sambrook et al., *supra*. See, e.g., Chapter 11 and pages 11.31-11.32 and 11.40-11.44, which describes radiolabeling of short probes, and pages 11.45-11.53, which describes hybridization conditions for oligonucleotide probes, including specific conditions for probe hybridization (pages 11.50-11.51). Methods of performing PCR using primers are also well-known in the art. See, e.g., Sambrook et al., *supra* and Ausubel et al., *supra*. PCR methods may be used to identify and/or isolate allelic variants and fragments of the nucleic acid molecules of the invention; PCR may also be used to identify and/or isolate nucleic acid molecules that hybridize to the primers and that may be amplified, and may be used to isolate nucleic acid molecules that encode homologous proteins, analogs, fusion protein or muteins of the invention.

Methods of Using Thioesterases for Biosynthesis of Compounds – Manipulations of *Dpt* Genes

Genes of the daptomycin biosynthetic gene cluster of the invention may be manipulated in a variety of ways to produce new biosynthetic peptide products or to alter the regulation of one or more genes expressed from the gene cluster. See, e.g., Figure 1.

Disruption of a Gene Encoding a Thioesterase

In one aspect, the invention provides a method of disrupting or deleting a gene encoding a thioesterase that is involved in a NRPS or PKS pathway in a bacterial cell. Preferably, the method comprises the step of disrupting or deleting a gene or portion thereof that encodes a thioesterase in a daptomycin biosynthetic gene cluster. Disruption or deletion of a gene encoding an integral thioesterase would be likely to result in the production of compounds that are intermediates to the final product. In one aspect, a gene or portion thereof encoding an integral thioesterase may be disrupted or deleted. In a preferred embodiment, disruption or deletion of a gene encoding an integral thioesterase of the daptomycin biosynthetic gene cluster in *S. roseosporus* would produce a linear lipopeptide compound. The linear lipopeptide compound may be used directly if its release from the NRPS were to be catalyzed by a different endogenous or exogenously provided thioesterase activity within the host cell. Such linear lipopeptide compounds, if not released from the NRPS by an endogenous thioesterase activity, may be useful intermediates for testing potential but as yet unidentified thioesterase polypeptides or for testing thioesterase fusion, fragment, mutein, derivative, analog or homolog polypeptides for activity. The linear lipopeptide compound may alternatively be used as an intermediate for production of novel lipopeptides.

In another aspect, a gene encoding a free thioesterase may be disrupted or deleted in a bacterial cell comprising an NRPS. Because free thioesterases are thought to be involved in proofreading of the peptide compounds produced in NRPS, disruption or deletion of a gene encoding a free thioesterase leads to the production of compounds that have mutations compared to the compound produced in the presence of the free thioesterase. These mutated compounds may be used to generate novel lipopeptides. See, e.g., Example 16.

In a preferred embodiment, the method comprises the step of disrupting or deleting the thioesterase-encoding portion of *dptD* or disrupting or deleting *dptH* in a daptomycin biosynthetic gene cluster. In an even more preferred embodiment, the method comprises the step of disrupting or deleting a gene encoding a thioesterase having an amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or having the amino acid sequence of SEQ ID NO: 8. The invention also comprises a method of disrupting or deleting a gene encoding a thioesterase wherein the gene is one that selectively hybridizes or is homologous to a gene encoding a thioesterase having an amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or the amino acid sequence of SEQ ID NO: 8. In another preferred embodiment, disruption or deletion of a thioesterase may be combined with the methods of altering the gene cluster involved in non-ribosomal peptide synthesis, as described below.

Disruption of a gene encoding a thioesterase may be accomplished by any method known to one having ordinary skill in the art following the teachings of the instant specification. In a preferred embodiment, disruption of a gene encoding a thioesterase may be accomplished by targeted gene disruption using methods taught, e.g., in Hosted and Baltz, J. Bacteriol., 179, pp. 180-186 (1997); Butler et al., Chem. Biol., 6, pp. 287-292 (1999); and Xue et al., Proc. Natl. Acad. Sci. U.S.A., 95, pp. 12111-12116 (1998), each of which is incorporated herein by reference in its entirety. See, e.g., Example 11.

Alteration of Site of Cyclization and Cyclic Peptide Produced Using Thioesterases

In a naturally-occurring polypeptide involved in NRPS, an integral thioesterase is located at the carboxy-terminus of the polypeptide, where it is involved in product cyclization. In one aspect, the invention provides a method to alter the site of cyclization of a cyclic peptide (or release of a linear peptide) by changing the location of a module encoding a thioesterase. In one embodiment, the site of cyclization may be altered by inserting the module encoding the thioesterase into the gene encoding the polypeptide involved in NRPS in a region that is upstream of the region in which the thioesterase module naturally occurs. In this embodiment, the cyclic peptide that is produced will be smaller than the naturally-occurring cyclic peptide. See, e.g., Example 12.

In a preferred embodiment, the module encodes an integral thioesterase from a daptomycin biosynthetic gene cluster. In a more preferred embodiment, the module comprises the thioesterase domain of DptD. In an even more preferred embodiment, the module encodes a polypeptide having all or a portion of the amino acid sequence of SEQ ID NO: 7, preferably a portion of SEQ ID NO: 7 that comprises the thioesterase domain. In another preferred embodiment, the module comprises a nucleic acid molecule that is homologous to or selectively hybridizes to a nucleic acid molecule encoding all or a portion of the thioesterase domain of SEQ ID NO: 7 or to a nucleic acid molecule encoding the thioesterase domain that comprises all or a portion of the nucleic acid sequence of SEQ ID NO: 3.

Alternatively, other modules that are involved in adding amino acids to the peptide (or otherwise modifying amino acids within the peptide) may be inserted upstream of the module encoding the thioesterase. See, e.g., Example 12. Such modules include a minimal module comprising at least an adenylation domain and a thiolation or acyl carrier domain. In a preferred embodiment, the inserted module would also include a condensation domain. Additional domains may also be inserted upstream of the thioesterase module including an M domain, an E domain and/or a Cy domain. The type of module(s) that would be inserted upstream of the thioesterase domain would depend upon the type of amino acid residues that were desired. Methods of inserting modules that will add and/or modify a specific amino acid are well known in the art. See, e.g., Mootz et al., Current Opinion in Biotechnology, 10, pp. 341-348 (1999), herein incorporated by reference in its entirety. Addition of one or more modules upstream of the thioesterase will produce a polypeptide involved in NRPS that is capable of synthesizing a cyclic peptide that is larger and that may contain different amino acid residues than the naturally-occurring cyclic peptide.

In vitro Use of Thioesterases for Production of Linear And Cyclic Peptides

In another aspect, the thioesterases of the invention may be used for production of cyclic peptides *in vitro*. See, e.g., Example 13. This method is particularly useful for generating novel linear and cyclic peptides by generating the peptide-compound substrate *in vitro*, e.g., by peptide synthesis and chemical linkage to a compound, and then cyclizing the peptide (or releasing a linear peptide) with an isolated thioesterase. In one embodiment, a thioesterase of the invention is recombinantly produced or is isolated from bacteria. The thioesterase of the invention is then incubated with a compound that can act as a substrate for the thioesterase. In a preferred embodiment, the thioesterase is incubated with a peptide of interest chemically linked to a compound. The peptide-compound substrate is one that is recognized by the thioesterase. In a preferred embodiment, the peptide-compound substrate is peptide-N-acetylcysteamine (NAC) thioester (peptide-SNAC). See, e.g., Trauger et al., Nature, 407, pp. 215-218 (2000). In another preferred embodiment, the peptide-compound substrate is peptide-pantetheine thioester. In another preferred embodiment, the peptide-compound substrate is a peptide thioester where the thiol is a suitable pantetheine mimic. One may use these methods for drug discovery programs using high throughput screening. See, e.g., Example 14. One having ordinary skill in the art in light of the teachings of the instant specification realize that not all peptide-compound substrates will be cyclized and/or released with the same efficiency as a peptide-compound substrate wherein the peptide has a sequence that is the same as the naturally-occurring peptide of daptomycin. Certain alterations in the peptide sequence, compared to the naturally-occurring sequence, are likely to decrease the rate of cyclization by the thioesterase. In particular, alterations of the first, penultimate and ultimate amino acids are likely to decrease the rate of cyclization. See, e.g., Trauger et al., *Nature* 407:215-218 (2000).

The peptide-compound substrate is incubated with the thioesterase under conditions in which the thioesterase can cyclize and/or release the peptide. In a preferred embodiment, the thioesterase is one that is derived from a daptomycin biosynthetic gene cluster. In a more preferred embodiment, the thioesterase is encoded by the thioesterase-encoding domain of *dptD* or by *dptH*. More preferably, the thioesterase has an amino acid sequence of the thioesterase domain of SEQ ID NO: 7

or of SEQ ID NO: 8, or is a homologous protein, fusion protein, mutein, analog, derivative or fragment thereof having thioesterase activity.

In Vivo Use of Thioesterases

Another use of the genes of the present invention is to improve the yield of a product in a cell expressing an NRPS. See, e.g., Example 11. Nucleic acid molecules that may be used to increase yield include nucleic acid molecules that encode positive regulatory factors, acyl CoA thioesterase, ABC transporters, NovABC-related polypeptides, DptA, DptB, DptC, or DptD, polypeptides that encode daptomycin resistance and daptomycin thioesterases, including DptD and DptH. The complete daptomycin biosynthetic gene cluster, daptomycin NRPS or any domain or subunit thereof may also be duplicated. In a preferred embodiment, a free and/or an integral thioesterase from a daptomycin biosynthetic gene cluster are introduced into a cell to improve production of daptomycin. In another preferred embodiment, the additional copies of a thioesterase may be introduced into a cell comprising altered NRPS polypeptides, as described *supra*. Without wishing to be bound by any theory, additional copies of a free and/or an integral thioesterase may improve the NRPS processing of the peptide by increasing the proofreading capacity (e.g., the free thioesterase) or the cyclization and/or peptide release capacity (e.g., the integral thioesterase) of the bacterial cell.

In a preferred embodiment, additional copies of a nucleic acid molecule encoding thioesterase may be introduced into a cell. See, e.g., Example 11. Introduction of the thioesterase may be performed by any method known in the art. In a more preferred embodiment, the additional copies of the gene are under the regulatory control of strong expression control sequences. These sequences may be derived from another thioesterase gene or may be derived from heterologous sequences, as described *supra*. Further, a nucleic acid molecule encoding a thioesterase may be introduced into a cell such that it is expressed as a separate polypeptide. This may be especially useful for a free thioesterase. Alternatively, a nucleic acid molecule encoding a thioesterase may be introduced into a cell such that it forms part of a multi-domain protein. This can be accomplished, e.g., by homologous recombination into a polypeptide which forms or interacts with an NRPS. This may be especially useful, although not required, for an integral thioesterase.

In another embodiment, copies of a free and/or an integral thioesterase may be introduced into a cell that expresses a NRPS complex that is other than a daptomycin biosynthetic gene cluster. See, e.g., Example 16. In one preferred embodiment, the complex is a NRPS complex. In another preferred embodiment, the complex is a PKS complex or a mixed PKS/NRPS complex. Numerous PKS and NRPS complexes are known in the art. See, e.g., complexes that produce vancomycin, bleomycin, A54145, CDA, amphomycin, echinocandin, cyclosporin, erythromycin, tylosin, monensin, avermectin, penicillin, cephalosporin, pristinamycins, erythromycin, rapamycin, spinosyn, didemnin, discobahamian, and epothilone. As described above, addition of a free and/or an integral thioesterase may improve the NRPS or PKS processing of a peptide by increasing the proofreading capacity (the free thioesterase) or the cyclization capacity (the integral thioesterase) of the bacterial cell. Addition of a free and/or integral thioesterase may be achieved by the methods described above.

In a preferred embodiment, a nucleic acid molecule encoding a thioesterase that is introduced into a cell is a thioesterase from a daptomycin biosynthetic gene cluster. In a preferred embodiment, the gene is the thioesterase-encoding domain of *dptD* or is *dptH*. More preferably, the nucleic acid molecule encodes a thioesterase having an amino acid sequence of the thioesterase domain of SEQ ID NO: 7 or SEQ ID NO: 8, or is a homologous protein, fusion protein, mutein, derivative, analog or fragment thereof having thioesterase activity.

Methods of Altering Gene Clusters for Production of Novel Compounds by NRPS

Alteration of NRPS Polypeptide Modules and Domains

In another aspect, the invention provides a method of altering the number or position of the modules in an NRPS. In one embodiment, one or more modules may be deleted from the NRPS. These deletions will result in synthesis by the NRPS of a peptide product that is shorter than the naturally-occurring one. In another embodiment, one or more modules or domains may be added to the NRPS. In this case, the peptide synthesized by the NRPS will be longer than the naturally-occurring one or will have an additional chemical change, e.g., if an epimerization domain or a methylation domain is added, the resultant peptide will contain an extra D-amino acid or will contain a methylated amino acid, respectively. In a yet further embodiment, one or more modules may be mutated, e.g., an adenylation domain may be mutated such that it has a different amino acid specificity than the naturally-occurring adenylation domain. The amino acid pocket code for the daptomycin NRPS – which determines which amino acid will bind within each adenylation domain of modules 1-13 – is described in Example 5; see also Table 2. With the amino acid code in hand, one of skill in the art can perform mutagenesis, by a variety of well known techniques, to exchange the code in one module for another code, thus altering the ultimate amino acid composition and/or sequence of the resulting peptide synthesized by the altered NRPS. See, e.g., Example 12A.

In a still further embodiment, one or more modules or domains may be substituted with another module or domain. In this case, the peptide produced by the altered NRPS will have, e.g., one or more different amino acids compared to the naturally-occurring peptide. In addition, different combinations of insertions, deletions, substitutions and mutations may be used to produce a peptide of interest. Further, the invention contemplates these altered NRPS complexes with and without an integral thioesterase domain. See, e.g., Example 12B-J.

The peptides produced by the NRPSs may be useful as new compounds or may be useful in producing new compounds. In a preferred embodiment, the new compounds are useful as or may be used to produce antibiotic compounds. In another preferred embodiment, the new compounds are useful as or may be used to produce other peptides having useful activities, including but not limited to antibiotic, antifungal, antiviral, antiparasitic, antimitotic, cytostatic, antitumor, immunomodulatory, anti-cholesterolemic, siderophore, agrochemical (e.g., insecticidal) or physicochemical (e.g., surfactant) properties. In a more preferred embodiment, the compounds produced using an altered NRPS polypeptide may be used in the synthesis of daptomycin-related compounds, including those described in United States Application Nos. 09/738,742, 09/737,908 and 09/739,535, filed December 15, 2000.

In addition, diverse variants of non-ribosomally synthesized peptides and polyketides may be achieved by altering the pools of available substrates during host cell cultivation. Commercial production of daptomycin, for example, is the result of cultivating the daptomycin producer *Streptomyces roseosporus* in the presence of decanoic acid, which alters the lipopeptide profile of the final products. See, e.g., United States Patent 4,885,243. The feeding of N-acetyl cysteamine (SNAC) analogs of polyketide intermediates resulted in substantial increases in incorporation of the intermediates into the polyketide, when compared to the free carboxylic acid or ester analogs. See, e.g., S. Yue et al., J. Am. Chem. Soc., 109, pp. 1253-1255 (1987); D.E. Cane and C-C Yang, J. Am. Chem. Soc., 109, 1255-1257 (1987); D.E. Cane et al., J. Am. Chem. Soc., 115, pp. 522-526 and 527-535 (1993); D.E. Cane et al., J. Am. Chem. Soc., 117, pp. 633-634 (1995); R. Pieder et al., J. Am. Chem. Soc., 117, pp. 11373-11374 (1995); each of which is incorporated herein by reference in its entirety. SNAC analogs of amino acids have been incorporated into a NRPS *in vitro*. D.E. Ehmann et al., Chem. Biol., 7, pp. 765-772 (2000). Thus it should be possible to feed SNAC or other pantetheine mimics to incorporate unnatural substrates into a NRPS-produced peptide.

Further diversity of non-ribosomally synthesized peptides and polyketides may also be achieved by expressing one or more NRPS and PKS genes (encoding natural, hybrid or otherwise altered modules or domains) in heterologous host cells, i.e., in host cells other than those from which the NRPS and PKS genes or modules originated.

In addition, one may express an ABC transporter or other polypeptide involved in antibiotic resistance in order to increase the resistance of a bacterial cell to daptomycin or a related compound. The ABC transporter may be overexpressed in a autologous cell (i.e., a cell that comprises the gene) or may be expressed in a heterologous cell (i.e., a cell that normally does not have the gene). Further, one may express an ABC transporter gene of the invention or another polypeptide involved in antibiotic resistance described herein in order to be able to select cells that are resistant to daptomycin. This selection may be useful for determining mechanisms of daptomycin resistance or may be used in standard molecular biological techniques in which antibody resistance is selected for.

Compounds Of The Invention, Pharmaceutical Compositions Thereof And Methods Of Treating Using Compounds And Compositions

Another object of the instant invention is to provide peptides or lipopeptides that may be produced by using the thioesterases, an NRPS or subunits thereof of the instant invention, as well as salts, esters, amides, ethers and protected forms thereof, and pharmaceutical formulations comprising these peptides, lipopeptides or their salts. In a preferred embodiment, the lipopeptide is daptomycin or a daptomycin-related lipopeptide, as described *supra*.

One may determine whether a peptide, lipopeptide or other compound of this invention has antibiotic activity using any of a variety of routine and well-known protocols in the art. One may use either an isolated or purified compound or may use an unpurified compound that is present in, e.g., fermentation culture broth or in a cell lysate. One may use either or both a gram-positive or a gram-negative bacterial test strain, and may use a variety of test strains to determine efficacy. In a preferred embodiment, the bacterial test strain will be a gram-positive test strain. In a more preferred embodiment, the bacterial test strain will be a *Staphylococcus*, more preferably *S. aureus*. An example of methods that can be used to determine antibiotic activity are provided in United States Patents 4,208,408 and 4,537,717. One having ordinary skill in the art will recognize that other potential antibiotics and other test strains may be used.

Peptides, lipopeptides or pharmaceutically acceptable salts thereof can be formulated for oral, intravenous, intramuscular, subcutaneous, aerosol, topical or parenteral administration for the therapeutic or prophylactic treatment of diseases, particularly bacterial infections. In a preferred embodiment, the lipopeptide is daptomycin or a daptomycin-related lipopeptide. Reference herein to "daptomycin," "daptomycin-related lipopeptide" or "lipopeptide" includes pharmaceutically acceptable salts thereof. Peptides, including daptomycin or daptomycin-related lipopeptides, can be formulated using any pharmaceutically acceptable carrier or excipient that is compatible with the peptide or with the lipopeptide of interest. See, e.g., Handbook of Pharmaceutical Additives: An International Guide to More than 6000 Products by Trade Name, Chemical, Function, and Manufacturer, Ashgate

Publishing Co., eds., M. Ash and I. Ash, 1996; The Merck Index: An Encyclopedia of Chemicals, Drugs and Biologicals, ed. S. Budavari, annual; Remington's Pharmaceutical Sciences, Mack Publishing Company, Easton, PA; Martindale: The Complete Drug Reference, ed. K. Parfitt, 1999; and Goodman & Gilman's The Pharmaceutical Basis of Therapeutics, Pergamon Press, New York, NY, ed. L. S. Goodman et al.; the contents of which are incorporated herein by reference, for a general description of the methods for administering various antimicrobial agents for human therapy. Peptides or lipopeptides of this invention can be mixed with conventional pharmaceutical carriers and excipients and used in the form of tablets, capsules, elixirs, suspensions, syrups, wafers, creams and the like. Peptides or lipopeptides may be mixed with other therapeutic agents and antibiotics, such as discussed herein. The compositions comprising a compound of this invention will contain from about 0.1 to about 90% by weight of the active compound, and more generally from about 10 to about 30%.

The compositions of the invention can be delivered using controlled (e.g., capsules) or sustained release delivery systems (e.g., bioerodable matrices). Exemplary delayed release delivery systems for drug delivery that are suitable for administration of the compositions of the invention are described in U.S. Patent Nos. 4,452,775 (issued to Kent), 5,239,660 (issued to Leonard), 3,854,480 (issued to Zaffaroni).

The compositions may contain common carriers and excipients, such as corn starch or gelatin, lactose, sucrose, microcrystalline cellulose, kaolin, mannitol, dicalcium phosphate, sodium chloride and alginic acid. The compositions may contain croscarmellose sodium, microcrystalline cellulose, corn starch, sodium starch glycolate and alginic acid.

Tablet binders that can be included are acacia, methylcellulose, sodium carboxymethylcellulose, polyvinylpyrrolidone (Povidone), hydroxypropyl methylcellulose, sucrose, starch and ethylcellulose.

Lubricants that can be used include magnesium stearate or other metallic stearates, stearic acid, silicone fluid, talc, waxes, oils and colloidal silica.

Flavoring agents such as peppermint, oil of wintergreen, cherry flavoring or the like can also be used. It may also be desirable to add a coloring agent to make the dosage form more aesthetic in appearance or to help identify the product.

For oral use, solid formulations such as tablets and capsules are particularly useful. Sustained release or enterically coated preparations may also be devised. For pediatric and geriatric applications, suspensions, syrups and chewable tablets are especially suitable. For oral administration, the pharmaceutical compositions are in the form of, for example, a tablet, capsule, suspension or liquid. The pharmaceutical composition is preferably made in the form of a dosage unit containing a therapeutically-effective amount of the active ingredient. Examples of such dosage units are tablets and capsules. For therapeutic purposes, the tablets and capsules which can contain, in addition to the active ingredient, conventional carriers such as binding agents, for example, acacia gum, gelatin, polyvinylpyrrolidone, sorbitol, or tragacanth; fillers, for example, calcium phosphate, glycine, lactose, maize-starch, sorbitol, or sucrose; lubricants, for example, magnesium stearate, polyethylene glycol, silica, or talc; disintegrants, for example, potato starch, flavoring or coloring agents, or acceptable wetting agents. Oral liquid preparations generally are in the form of aqueous or oily solutions, suspensions, emulsions, syrups or elixirs may contain conventional additives such as suspending agents, emulsifying agents, non-aqueous agents, preservatives, coloring agents and flavoring agents. Oral liquid preparations may comprise lipopeptide micelles or monomeric forms of the lipopeptide. Examples of additives for liquid preparations include acacia, almond oil, ethyl alcohol, fractionated coconut oil, gelatin, glucose syrup, glycerin, hydrogenated edible fats, lecithin, methyl cellulose, methyl or propyl *para*-hydroxybenzoate, propylene glycol, sorbitol, or sorbic acid.

For intravenous (IV) use, a water soluble form of the peptide or lipopeptide can be dissolved in any of the commonly used intravenous fluids and administered by infusion. Intravenous formulations may include carriers, excipients or stabilizers including, without limitation, calcium, human serum albumin, citrate, acetate, calcium chloride, carbonate, and other salts. Intravenous fluids include, without limitation, physiological saline or Ringer's solution. Peptides or lipopeptides also may be placed in injectors, cannulae, catheters and lines.

Formulations for parenteral administration can be in the form of aqueous or non-aqueous isotonic sterile injection solutions or suspensions. These solutions or suspensions can be prepared from sterile powders or granules having one or more of the carriers mentioned for use in the formulations for oral administration. Lipopeptide micelles may be particularly desirable for parenteral administration. The compounds can be dissolved in polyethylene glycol, propylene glycol, ethanol, corn oil, benzyl alcohol, sodium chloride, and/or various buffers. For intramuscular preparations, a sterile formulation of a lipopeptide compound or a suitable soluble salt form of the compound, for example the hydrochloride salt, can be dissolved and administered in a pharmaceutical diluent such as Water-for-Injection (WFI), physiological saline or 5% glucose.

Injectable depot forms may be made by forming microencapsulated matrices of the compound in biodegradable polymers such as polylactide-polyglycolide. Depending upon the ratio of drug to polymer and the nature of the particular polymer employed, the rate of drug release can be controlled. Examples of other biodegradable polymers include poly(orthoesters) and poly(anhydrides). Depot injectable formulations are also prepared by entrapping the drug in microemulsions that are compatible with body tissues.

For topical use the compounds of the present invention can also be prepared in suitable forms to be applied to the skin, or mucus membranes of the nose and throat, and can take the form of creams, ointments, liquid sprays or inhalants, lozenges, or throat paints. Such topical formulations further can include chemical compounds such as dimethylsulfoxide (DMSO) to facilitate surface penetration of the active ingredient. For topical preparations, a sterile formulation of daptomycin, daptomycin-related lipopeptide or suitable salt forms thereof, may be administered in a cream, ointment, spray or other topical dressing. Topical preparations may also be in the form of bandages that have been impregnated with daptomycin or a daptomycin-related lipopeptide composition.

For application to the eyes or ears, the compounds of the present invention can be presented in liquid or semi-liquid form formulated in hydrophobic or hydrophilic bases as ointments, creams, lotions, paints or powders.

For rectal administration the compounds of the present invention can be administered in the form of suppositories admixed with conventional carriers such as cocoa butter, wax or other glyceride.

For aerosol preparations, a sterile formulation of the peptide or lipopeptide or salt form of the compound may be used in inhalers, such as metered dose inhalers, and nebulizers. A sterile formulation of a lipopeptide micelle may also be used for aerosol preparation. Aerosolized forms may be especially useful for treating respiratory infections, such as pneumonia and sinus-based infections.

Alternatively, the compounds of the present invention can be in powder form for reconstitution in the appropriate pharmaceutically acceptable carrier at the time of delivery. In one embodiment, the unit dosage form of the compound can be a solution of the compound or a salt thereof, in a suitable diluent in sterile, hermetically sealed ampules. The concentration of the compound in the unit dosage may vary, e.g. from about 1 percent to about 50 percent, depending on the compound used and its solubility and the dose desired by the physician. If the compositions contain dosage units, each dosage unit preferably contains from 50-500 mg of the active material. For adult human treatment, the dosage employed preferably ranges from 100 mg to 3 g, per day, depending on the route and frequency of administration.

In a further aspect, this invention provides a method for treating an infection, especially those caused by gram-positive bacteria, in humans and other animals. The term "treating" is used to denote both the prevention of an infection and the control of an established infection after the host animal has become infected. An established infection may be one that is acute or chronic. The method comprises administering to the human or other animal an effective dose of a compound of this invention. An effective dose of daptomycin, for example, is generally between about 0.1 and about 25 mg/kg daptomycin, daptomycin-related lipopeptide or pharmaceutically acceptable salts thereof. The daptomycin or daptomycin-related lipopeptide may be monomeric or may be part of a lipopeptide micelle. A preferred dose is from about 1 to about 25 mg/kg of daptomycin or daptomycin-related lipopeptide or pharmaceutically acceptable salts thereof. A more preferred dose is from about 1 to 12 mg/kg daptomycin or a pharmaceutically acceptable salt thereof. These dosages for daptomycin may be used as a starting point by one of skill in the art to determine and optimize effective dosages of other linear and cyclic peptides produced by the modified NRPS complexes of the invention.

In one embodiment, the invention provides a method for treating an infection, especially those caused by gram-positive bacteria, in a subject with a therapeutically-effective amount of modified daptomycin or other antibacterial peptide or lipopeptide produced by a modified NRPS of the invention. The daptomycin or antibacterial peptide or lipopeptide may be monomeric or in a lipopeptide micelle. Exemplary procedures for delivering an antibacterial agent are described in U.S. Patent No. 5,041,567, issued to Rogers and in PCT patent application number EP94/02552 (publication no. WO 95/05384), the entire contents of which documents are incorporated in their entirety herein by reference. As used herein the phrase "therapeutically-effective amount" means an amount of modified daptomycin or other antibacterial peptide or lipopeptide produced by a modified NRPS according to the present invention, that prevents the onset, alleviates the symptoms, or stops the progression of a bacterial infection. The term "treating" is defined as administering, to a subject, a therapeutically-effective amount of a compound of the invention, both to prevent the occurrence of an infection and to control or eliminate an infection. The term "subject", as described herein, is defined as a mammal, a plant or a cell culture. In a preferred embodiment, a subject is a human or other animal patient in need of peptide or lipopeptide compound treatment.

The peptide or lipopeptide antibiotic compound can be administered as a single daily dose or in multiple doses per day. The treatment regime may require administration over extended periods of time, e.g., for several days or for from two to four weeks. The amount per administered dose or the total amount administered will depend on such factors as the nature and severity of the infection, the age and general health of the patient, the tolerance of the patient to the antibiotic and the microorganism or microorganisms involved in the infection. A method of administration is disclosed in United States Serial No. 09/406,568, filed September 24, 1999, herein incorporated by reference, which claims the benefit of U.S. Provisional Application Nos. 60/101,828, filed September 25, 1998, and 60/125,750, filed March 24, 1999.

The methods of the present invention comprise administering modified daptomycin or other peptide or lipopeptide antibiotics, or pharmaceutical compositions thereof to a patient in need thereof in an amount that is efficacious in reducing or eliminating the gram-positive bacterial infection. The antibiotic may be administered orally, parenterally, by inhalation, topically, rectally, nasally, buccally, vaginally, or by an implanted reservoir, external pump or catheter. The antibiotic may be prepared for opthalmic or aerosolized uses. Modified daptomycin, a peptide or lipopeptide antibiotic produced by a modified NRPS of the invention, or a pharmaceutical compositions thereof, also may be directly injected or administered into an abscess, ventricle or joint. Parenteral administration includes subcutaneous, intravenous, intramuscular, intra-articular, intra-synovial, cisternal, intrathecal, intrahepatic, intralesional and intracranial injection or infusion. In a preferred embodiment, daptomycin or another peptide or lipopeptide is administered intravenously, subcutaneously or orally.

The method of the instant invention may be used to treat a patient having a bacterial infection in which the infection is caused or exacerbated by any type of gram-positive bacteria. In a preferred embodiment, modified daptomycin, daptomycin-related lipopeptide, or another peptide or lipopeptide antibiotic produced by a modified NRPS of the invention, or pharmaceutical compositions thereof, are administered to a patient according to the methods of this invention. In another preferred embodiment, the bacterial infection may be caused or exacerbated by bacteria including, but not limited to, methicillin-susceptible and methicillin-resistant staphylococci (including *Staphylococcus aureus, Staphylococcus epidermidis, Staphylococcus haemolyticus, Staphylococcus hominis, Staphylococcus saprophyticus,* and coagulase-negative staphylococci), glycopeptide intermediary- susceptible *Staphylococcus aureus* (GISA), penicillin-susceptible and penicillin-resistant streptococci (including *Streptococcus pneumoniae, Streptococcus pyogenes, Streptococcus agalactiae, Streptococcus* *avium, Streptococcus bovis, Streptococcus lactis, Streptococcus sangius* and *Streptococci* Group C, *Streptococci* Group G and viridans streptococci), enterococci (including vancomycin-susceptible and vancomycin-resistant strains such as *Enterococcus faecalis* and *Enterococcus faecium*), *Clostridium difficile*, *Clostridium clostridiiforme*, *Clostridium innocuum*, *Clostridium perfringens*, *Clostridium ramosum, Haemophilus influenzae, Listeria monocytogenes, Corynebacterium jeikeium*, *Bifidobacterium* spp., *Eubacterium aerofaciens, Eubacterium lentum, Lactobacillus acidophilus, Lactobacillus casei, Lactobacilllus plantarum, Lactococcus* spp., *Leuconostoc* spp., *Pediococcus, Peptostreptococcus anaerobius, Peptostreptococcus asaccarolyticus, Peptostreptococcus magnus, Peptostreptococcus micros, Peptostreptococcus prevotii, Peptostreptococcus productus, Propionibacterium acnes,* and *Actinomyces* spp.

The antibacterial activity of daptomycin against classically "resistant" strains is comparable to that against classically "susceptible" strains in *in vitro* experiments. In addition, the minimum inhibitory concentration (MIC) value for daptomycin against susceptible strains is typically 4-fold lower than that of vancomycin. Thus, in a preferred embodiment, modified daptomycin, daptomycin-related lipopeptide antibiotic, a peptide or lipopeptide antibiotic produced by the modified NRPS of the invention, or pharmaceutical compositions thereof, are administered according to the methods of this invention to a patient who exhibits a bacterial infection that is resistant to other antibiotics, including vancomycin. In addition, unlike glycopeptide antibiotics, daptomycin exhibits rapid, concentration-dependent bactericidal activity against gram-positive organisms. Thus, in a preferred embodiment, daptomycin, a lipopeptide antibiotic, or pharmaceutical compositions thereof are administered according to the methods of this invention to a patient in need of rapidly acting antibiotic therapy.

The method of the instant invention may be used for a gram-positive bacterial infection of any organ or tissue in the body. These organs or tissue include, without limitation, skeletal muscle, skin, bloodstream, kidneys, heart, lung and bone. The method of the invention may be used to treat, without limitation, skin and soft tissue infections, bacteremia and urinary tract infections. The method of the invention may be used to treat community acquired respiratory infections, including, without limitation, otitis media, sinusitis, chronic bronchitis and pneumonia, including pneumonia caused by drug-resistant *Streptoococcus pneumoniae* or *Haemophilus influenzae*. The method of the invention also may be used to treat mixed infections that comprise different types of gram-positive bacteria, or which comprise both gram-positive and gram-negative bacteria, including aerobic, caprophilic or anaerobic bacteria. These types of infections include intra-abdominal infections and obstetrical/gynecological infections. The methods of the invention may be used in step-down therapy for hospital infections, including, without limitation, pneumonia, intra-abdominal sepsis, skin and soft tissue infections and bone and joint infections. The method of the invention also may be used to treat an infection including, without limitation, endocarditis, nephritis, septic arthritis and osteomyelitis. In a preferred embodiment, any of the above-described diseases may be treated using daptomycin, lipopeptide antibiotic, or pharmaceutical compositions thereof. Further, the diseases may be treated using daptomycin or lipopeptide antibiotic in either a monomeric or micellar form.

Modified daptomycin, daptomycin-related lipopeptide, or another peptide or lipopeptide produced by a modified NRPS according to the invention, may also be administered in the diet or feed of a patient or animal. If administered as part of a total dietary intake, the amount of modified daptomycin or other peptide or lipopeptide can be less than 1% by weight of the diet and preferably no more than 0.5% by weight. The diet for animals can be normal foodstuffs to which modified daptomycin or the other peptide or lipopeptide can be added or it can be added to a premix.

The method of the instant invention may also be practiced while concurrently administering one or more antifungal agents and/or one or more antibiotics other than modified daptomycin or other peptide or lipopeptide antibiotic. Co-administration of an antifungal agent and an antibiotic other than modified daptomycin or another peptide or lipopeptide antibiotic may be useful for mixed infections such as those caused by different types of gram-positive bacteria, those caused by both gram-positive and gram-negative bacteria, or those that caused by both bacteria and fungus. Furthermore, modified daptomycin or other peptide or lipopeptide antibiotic may improve the toxicity profile of one or more co-administered antibiotics. It has been shown that administration of daptomycin and an aminoglycoside may ameliorate renal toxicity caused by the aminoglycoside. In a preferred embodiment, an antibiotic and/or antifungal agent may be administered concurrently with modified daptomycin, other peptide or lipopeptide antibiotic, or in pharmaceutical compositions comprising modified daptomycin or another peptide or lipopeptide antibiotic.

Antibacterial agents and classes thereof that may be co-administered with modified daptomycin or other peptide or lipopeptide antibiotics include, without limitation, penicillins and related drugs, carbapenems, cephalosporins and related drugs, aminoglycosides, bacitracin, gramicidin, mupirocin, chloramphenicol, thiamphenicol, fusidate sodium, lincomycin, clindamycin, macrolides, novobiocin, polymyxins, rifamycins, spectinomycin, tetracyclines, vancomycin, teicoplanin, streptogramins, anti-folate agents including sulfonamides, trimethoprim and its combinations and pyrimethamine, synthetic antibacterials including nitrofurans, methenamine mandelate and methenamine hippurate, nitroimidazoles, quinolones, fluoroquinolones, isoniazid, ethambutol, pyrazinamide, para-aminosalicylic acid (PAS), cycloserine, capreomycin, ethionamide, prothionamide, thiacetazone, viomycin, eveminomycin, glycopeptide, glycylcylcline, ketolides, oxazolidinone; imipenen, amikacin, netilmicin, fosfomycin, gentamicin, ceftriaxone, Ziracin, LY 333328, CL 331002, HMR 3647, Linezolid, Synercid, Aztreonam, and Metronidazole, Epiroprim, OCA-983, GV-143253, Sanfetrinem sodium, CS-834, Biapenem, A-99058.1, A-165600, A-179796, KA 159, Dynemicin A, DX8739, DU 6681; Cefluprenam, ER 35786, Cefoselis, Sanfetrinem celexetil, HGP-31, Cefpirome, HMR-3647, RU-59863, Mersacidin, KP 736, Rifalazil; Kosan, AM 1732, MEN 10700, Lenapenem, BO 2502A, NE-1530, PR 39, K130, OPC 20000, OPC 2045, Veneprim, PD 138312, PD 140248, CP 111905, Sulopenem, ritipenam acoxyl, RO-65-5788, Cyclothialidine, Sch-40832, SEP-132613, micacocidin A, SB-275833, SR-15402, SUN A0026, TOC 39, carumonam, Cefozopran, Cefetamet pivoxil, and T 3811.

In a preferred embodiment, antibacterial agents that may be co-administered with modified daptomycin or peptide or lipopeptide antibiotic produced by a modified NRPS according to this invention include, without limitation, imipenen, amikacin, netilmicin, fosfomycin, gentamicin, ceftriaxone, teicoplanin, Ziracin, LY 333328, CL 331002, HMR 3647, Linezolid, Synercid, Aztreonam, and Metronidazole.

Antifungal agents that may be co-administered with modified daptomycin or other peptide or lipopeptide antibiotic include, without limitation, Caspofungen, Voriconazole, Sertaconazole, IB-367, FK-463, LY-303366, Sch-56592, Sitafloxacin, DB-289 polyenes, such as Amphotericin, Nystatin, Primaricin; azoles, such as Fluconazole, Itraconazole, and Ketoconazole; allylamines, such as Naftifine and Terbinafine; and anti-metabolites such as Flucytosine. Other antifungal agents include without limitation, those disclosed in Fostel et al., Drug Discovery Today 5:25-32 (2000), herein incorporated by reference. Fostel et al. disclose antifungal compounds including Corynecandin, Mer-WF3010, Fusacandins, Artrichitin/LL 15G256γ, Sordarins, Cispentacin, Azoxybacillin, Aureobasidin and Khafrefungin.

Modified daptomycin or other peptide or lipopeptide antibiotics, including daptomycin-related lipopeptides, may be administered according to this method until the bacterial infection is eradicated or reduced. In one embodiment, modified daptomycin, or other peptide or lipopeptide produced by a modified NRPS according to the invention, is administered for a period of time from 3 days to 6 months. In a preferred embodiment, modified daptomycin, or other peptide or lipopeptide, is administered for 7 to 56 days. In a more preferred embodiment, modified daptomycin, or other peptide or lipopeptide is administered for 7 to 28 days. In an even more preferred embodiment, modified daptomycin or other peptide or lipopeptide antibiotic is administered for 7 to 14 days. In another embodiment, the antibiotic is administered for 3 to 7 days. Modified daptomycin, or other peptide or lipopeptide produced by a modified NRPS according to the invention, according to the invention may be administered for a longer or shorter time period if it is so desired.

In order that this invention may be more fully understood, the following examples are set forth. These examples are for the purpose of illustration only and are not to be construed as limiting the scope of the invention in any way.

EXAMPLE 1: Initial sequencing of the Streptomyces roseosporus daptomycin biosynthetic gene cluster

Streptomyces roseosporus strain A21978.6 (American Type Culture Collection Accession No. 31568) was used for the construction of a cosmid library. Genomic DNA was digested partially with *Sau*3A1 and alkaline phosphatase (Boehringer Mannheim Biochemicals). DNA of approximately 40 kb in length was isolated and ligated to *Bam*HI-digested cosmid pKC1471 and packaged with a Gigapack packaging extract (Stratagene, Inc.) as described in Hosted and Baltz, J. Bacteriol., 179, pp. 180-186 (1997). Packaged DNA was introduced into *E. coli* XL1-Blue-MFR' (Stratagene, Inc.) and individual clones containing cosmid DNA were stored as an ordered array in a 96-well dot blot apparatus. Twelve cultures from a row of microtiter wells were pooled, and screened by hybridization to a 2.1-kB *SphI* fragment of DNA from plasmid pRHB153 and to a 5.2-kB *DraI-KpnI* fragment from pRHB157, both containing NRPS sequences cloned from *S. roseosporus* (see McHenney et al., *supra*). Individual cosmids from the hybridizing pools were identified by hybridization to the same probes.

Cosmid and plasmid DNA was hydrodynamically sheared and then separated by electrophoresis on a standard 1% agarose gel. The separated DNA fragments 2500-3000 bp in length were excised from the gel and purified by the GeneClean™ procedure (BIO 101, Inc.). The ends of the gel-purified DNA fragments were then filled in or made blunt using T4 DNA polymerase. The DNA fragments were ligated to unique *Bst*XI-linker adapters (5'-GTCTTCACCACGGGG-3' – SEQ ID NO: , and 5'GTGGTGAAGAC-3' – SEQ ID NO: , in 100-1000 fold molar excess). These linkers are complementary to the *Bst*XI-cut pGTC vector (Genome Therapeutics Corp., Waltham, MA), while the overhang is not self-complementary. Therefore, the linkers will not concatemerize, nor will the open vector self-ligate easily. The linker-adapted inserts were separated from the unincorporated linkers by electrophoresis on a 1% agarose gel and purified using GeneClean™. The purified linker-adapted inserts were ligated to *Bst*XI-cut pGTC vector to construct "shotgun" subclone libraries.

The pGTC library was then transformed into DH5α competent cells (Gibco/BRL, DH5α transformation protocol). Transformation was assessed by plating onto antibiotic plates containing ampicillin and IPTG/Xgal (IPTG = isopropyl-b-D-thiogalactopyranoside; Xgal = 5-bromo-4-chloro-3-indoyl-b-D-thiogalactopyranoside.)

The plates were incubated overnight at 37°C. Transformants were plate purified and the purified clones containing the following plasmids were picked for further analysis.

Plasmids pRHB160, containing an insert of approximately 50 kb of *S. roseosporus* DNA, pRHB613, containing an insert of approximately 15 kb, pRHB614, containing an insert of approximately 13 kb, and pRHB159, containing an insert of approximately 51 kb, were chosen for DNA sequencing. (See McHenney, M.A. *et al., supra*).

Individual cultures of strains transformed with the above plasmids were grown overnight at 37°C. DNA was purified using a silica bead DNA preparation method (Engelstein, M. *et al., Microb. Comp. Genomics* 3(4):237-241, 1998). In this manner, 25 mg of DNA were obtained per clone. These purified DNA samples were then sequenced using primarily ABI dye-terminator chemistry. All subsequent steps were based on sequencing by ABI377 or Amersham automated DNA sequencing methods according to the manufacturer's instructions. The ABI dye terminator sequence reads were run on either ABI377 or Amersham MegaBace™ capillary machines. The data were transferred to UNIX machines following lane tracking of the gels. Base calls and quality scores were determined using the program PHRED (Ewing *et al., Genome Res. 8*:175-185, 1998). Reads were assembled using PHRAP (P. Green, Abstracts of DOE Human Genome Program Contractor-Grantee Workshop V, Jan. 1996, p.157) with default program parameters and quality scores. The initial assembly was done at 6x coverage.

EXAMPLE 2: Isolation and analysis of additional DNA molecules of the Streptomyces roseosporus biosynthetic gene cluster

Mycelium for preparation of megabase DNA was obtained from overnight cultures of *Streptomyces roseosporus* (NRRL11379) (ATCC No. 31568) shaken in F10A broth (2% agar, 25% soluble starch, 0.2% dextrose, 0.5% yeast extract, 0.5% peptone and 0.3% calcium carbonate) at 30°C. Washed cells were embedded in Seakem™ GTG agarose (FMC Bioproducts, 1% final concentration), incubated in lysozyme (2mg/mL TE) at 37°C for 3h, then lysed in 0.1x NLS + 0.2mg/mL proteinase K at 50°C overnight to release DNA into the gel matrix. Agarose containing DNA was washed with 1 mM EDTA (pH 8) before treatment with *Bam*HI at 37° C. Partially digested DNA was then subjected to a two-step size selection process in 0.6% agarose gels (in 0.5X TBE) by pulsed-field electrophoresis using a CHEF Mapper DRIII (Biorad) set at 6V/cm, 120° angle, 12°C. The first selection consisted of a 14 h run with a 22-44 sec linearly ramped switch time. Gel containing DNA co-migrating with 100-200 kb lambda concatamer size markers was excised and cast in a second gel for an 18 h run with a 3-5 sec linear ramp. DNA estimated at 75-145 kb relative to size markers was electroeluted (MiniProtean II Cell model, Biorad) in TAE.

The single-copy BAC library cloning vector pStreptoBAC V is derived from pBACe3.6 (Frengen, E., Weichenhan, D., Zhao, B., Osoegawa, K., van Geel, M. & de Jong, P.J., A modular, positive selection bacterial artificial chromosome vector with multiple cloning sites, Genomics, 58: 250-253 (1999)). The pBACe3.6 was modified to contain two markers, Amp^R for selection in *E. coli* and Apra^R for selection in *Streptomyces*, as well as oriT and attP sequences from the phage φC31 for conjugation and site specific integration in *Streptomyces*. See Figure 6. To prepare the pStreptoBAC V vector for ligation with the *S. roseosporus* DNA, the vector was first digested with *BamHI* and the reaction was inactivated by heat (65°C for 1h). DNA was then dephosporylated with Shrimp Alkaline Phosphatase for 30min. The two bands (13 kb and 3kb corresponding to the pUC fragment) were separated on 0.6% agarose gel and the 13 kb band was purified using Geneclean spin columns.

200 ng of the *S. roseosporus* DNA was ligated to 75 ng of *Bam*HI cut and phosphatased pStreptoBAC V vector DNA using 9 U of T4 DNA ligase (Promega) in a 150 μl reaction. After 16 h at 16°C, the ligations were heated at 65° C for 30 min, dialyzed against 10% polyethylene glycol 8000, and transformed into 10 μl of DH10B electrocompetent cells (Gibco/BRL) using a cell porator with voltage booster (Gibco/BRL) at 300 V and 4 kΩ. Cells were plated on media (LB agar) containing 100mg/mL apramycin and 5% sucrose. Analysis of sample clones showed a range of inserts from 39 kb to 105 kb. The mean insert size was 71.4 kb, with a standard deviation of 14.7 kb. Approximately 2,000 clones were archived at –80°C in 96-well microtiter plates.

This BAC library was screened using the polymerase chain reaction (PCR) using primer pairs P61/P62, P72/P73 and P74/P75, shown below. Nucleotide positions refer to the numbering of SEQ ID NO: 1, and "C" indicates that the primer sequence corresponds to the complementary strand of SEQ ID NO: 1:

Primer	Sequence	SEQ ID NO:	Nucleotide Position
P61	GCTCGTCCCCCTCCCCGCACT		41305-41325
P62	CGAACAGGTGGGCTTTGAGTGG		41993-42014 (C)
P72	CTTCGTGAACACCCTCGTCC		82104-82124
P73	GTTCGTCGAGGTCCAGTACG		83011-83030 (C)
P74	GCACCAGCGTGTGCGGATCG		92-111
P75	CACGTACGTGACGATCCTCG		799-818 (C)

PCR was performed under the following conditions: 94° C, 45 sec., 54° C, 30sec., 72° C, 1 min. for 32 cycles. Taq polymerase, as well as the accessory reagents, were supplied by Gibco BRL (Bethesda); all reactions included 5% DMSO.

Clone B12:03A05 was initially detected with primer pair P61/P62 (see above), and subsequently confirmed as a positive hit with the other two primer pairs. DNA of clone B12:03A05 was obtained by standard alkaline lysis procedures and used for DNA sequencing (see below).

A number of other clones that encompass parts of the daptomycin gene cluster (*dpt*-related clones) were isolated from the BAC library. These clones include 01G05 (insert size 82 kb), 06A12 (insert size 85 kb), 12F06 (insert size 65 kb), 18H04 (insert size 46 kb) and 20C09 (insert size 65 kb): See Figure 7, which shows a *Hin*DIII digest of these BAC clones. Other BACs that were isolated in the daptomycin gene cluster region include 09D02, 17F08, 05D08, 15H07, 21F10 and 16D12. These BACS cover 180 to 200 kb. Figure 8 shows the approximate location of the BAC clones relative to the daptomycin gene cluster.

Extension of the daptomycin biosynthetic gene cluster sequence determined in Example 2 was accomplished by sequencing 1 μg aliquots of BAC DNA from clone B12:03A05 using the ABI Prism Dye Terminator Cycle Sequencing Ready Reaction kit (Perkin Elmer), the manufacturer's recommended reaction mix and conditions, and the following primers (C indicates that the primer sequence corresponds to the complementary strand of SEQ ID NO: 1):

Primer	Sequence	SEQ ID NO:	Nucleotide Position
P76	CGTACTGGACCTCGACGACC		83011-83030
P78	CGACCAGCGTGTGTACGTCC		83611-83630
P92	AGTCCTCAGCCATCTCCTCG		84586-84605 (C)
P84	GAGACCGTCGGCGTGGACG		84224-84242
P95	AGGGCCACACCGTCGAACTCC		84711-84731
P86	ATCGTCGCCGACTACCTCGC		84797-84816
P96	GGCAGCTACCTCGTACTGG		85299-85317
P97	TGTACGACAGCGGCGTCGAAC		85961-85981
P101	CGATTCTCGGCATGTTCGCC		86638-86657
P105	TCGTCTCCTACATGACCTCG		87196-87215
P107	TTCACGGAAACCGAACGTCG		87866-87885
P111	GGTTCAGGCCGCAGCCAACG		88468-88487
P117	CGCTGACCTTGGTCAGAAGCC		89176-89196

Electrophanerograms were inspected and corrected as appropriate, and the sequences were aligned using the AssemblyLign Module of MacVector™. The aligned sequence (contig) was saved as a MacVector™ file for analysis and annotation. Identification of potential ORFs and potential stops/starts was performed using the open reading frames option in MacVector™.

Analysis of the 90kb sequence showed a total of 38 open reading frames in the daptomycin biosynthetic gene cluster region. See Figure 2. The ORFs range in size from 228 basepairs (bp) to 17.5 kb. The four largest ORFs are NRPS genes, as discussed below. One of the NRPS genes were predicted to have thioesterase activity based on the presence of conserved motifs, GXSXG (see Example 3). Another predicted open reading frames also encodes a protein with thioesterase activity (see Example 3). A number of potential ABC transporters were also identified.

The sequence of the daptomycin biosynthetic gene cluster is shown in SEQ ID NO: 1. See also Figure 2. The genes encoding the daptomycin non-ribosomal peptide synthetase (NRPS) are designated *dptA*, *dptB*, *dptC* and *dptD*. We designate as a promoter region all sequences upstream from the start of an ORF of interest that are not part of an upstream ORF. Because *dptA*, *dptB*, *dptC* and *dptD* have overlapping start and stop codons and apparently are translationally coupled (e.g., the TGA stop codon of *dptC* overlaps with the ATG start codon of *dptD*, which is associated with its own ribosome binding site), we thus indicate the promoter of the whole cluster (comprising *dptE*, *dptF*, *dptA*, *dptB*, *dptC* and *dptD*) as the daptomycin NPRS promoter.

The DNA sequence of the ORF of the daptomycin NRPS *dptA* gene (nucleotides 38555-56047 of SEQ ID NO: 1) is shown in SEQ ID NO: 10. The ORF is 17493 nucleotides in length. The amino acid sequence of the encoded DptA protein is shown in SEQ ID NO: 9. The protein is 5830 amino acid residues in length.

The DNA sequence of the ORF of the daptomycin NRPS *dptB* gene (nucleotides 56044-68361 of SEQ ID NO: 1) is shown in SEQ ID NO: 12. The ORF is 12318 nucleotides in length. The amino acid sequence of the encoded DptB protein is shown in SEQ ID NO: 11. The protein is 4105 amino acid residues in length.

The DNA sequence of the ORF of the daptomycin NRPS *dptC* gene (nucleotides 68358-78062 of SEQ ID NO: 1) is shown in SEQ ID NO: 14. The ORF is 9705 nucleotides in length. The amino acid sequence of the encoded DptC protein is shown in SEQ ID NO: 13. The protein is 3234 amino acid residues in length.

The DNA sequence of the ORF of the daptomycin NRPS *dptD* gene (nucleotides 78059-85198 of SEQ ID NO: 1) is shown in SEQ ID NO: 3. The ORF is 7140 nucleotides. The *dptD* gene ORF encodes a type I thioesterase (TEI) domain at the C-terminus. The amino acid sequence of the predicted DptD protein is shown in SEQ ID NO: 7 (see Figure 3). The protein is 2379 amino acids in length

The *dptE* and *dptF* are located between *dptA* and the daptomycin NPRS promoter.

The DNA sequence of the *dptH* thioesterase-encoding gene is shown in SEQ ID NO: 4 (nucleotides 85500-86352 of SEQ ID NO: 1); the promoter region of *dptH* is shown in SEQ ID NO: 5 (nucleotides 85500-85536 of SEQ ID NO: 1); and the open reading frame of *dptH* is shown in SEQ ID NO: 6 (nucleotides 85537-86352 of SEQ ID NO: 1). The amino acid sequence of the predicted DptH protein is shown in SEQ ID NO: 8 (see Figure 4).

The promoter region of the daptomycin NRPS (nucleotides 36018-36407 of SEQ ID NO: 1) is shown in SEQ ID NO: 2.

EXAMPLE 3: Identification of the dptD and dptH genes as thioesterases

Amino acid motifs typical of non-ribosomal peptide synthetases and thioesterases were identified by inspection of the *dptD* and *dptH* genes and predicted translation products thereof. The amino acid sequence motif GXSXG, wherein X is any one of the twenty L-amino acids that are inserted translationally into ribosomally produced proteins, is indicative of thioesterases (See Mootz, H.D., *et al.*, *J. Bacteriol.* 179:6843-6850, 1997, incorporated herein by reference in its entirety). SEQ ID NOs 7-8 were inspected for the GXSXG thioesterase motif. In SEQ ID NO:7, the amino acid sequence match to the thioesterase motif GWSFG was found at coordinates 2200-2204, encoded by nucleotides 84656-84670 of SEQ ID NO:1. In SEQ ID NO:8, the amino acid sequence match to the thioesterase motif GTSLG was found at coordinates 97-101, encoded by nucleotides 85825-85840 of SEQ ID NO:1.

The DptD protein of SEQ ID NO:7 was aligned to the CDA III protein of *Streptomyces coelicolor*. The alignment was performed using the Clustal W (v1.4) program in slow pairwise alignment mode. An open gap penalty of 10.0, an extend gap penalty of 0.1, and a blosum similarity matrix to the CDA III protein was used. The CDA III protein is a non-ribosomal peptide synthetase with a carboxy-terminal thioesterase domain (see GENBANK accession number AL035707, version AL035707.1 GI:4490978, hereby incorporated by reference in its entirety). The CDA III amino acid sequence used for the alignment was generated using the MacVector program by creating a contig from two GENBANK cosmid sequences, AL035707 and AL035640, and then translating the open reading frame in the contig annotated in GENBANK. The sequence comparison (Figure 3) revealed an alignment score of 7705 and 1223 conserved identities, indicating significant similarity between the two compared sequences. The GXSXG thioesterase motifs of the DptD protein and the CDA III protein were aligned in this analysis.

The GXSXG thioesterase motif of the DptH protein of SEQ ID NO: 8 was aligned to the GXSXG thioesterase motif of the CDA III protein of *Streptomyces coelicolor* (CAA71338 protein, see above). The alignment was performed the Clustal W (v1.4) program in slow pairwise alignment mode. An open gap penalty of 10.0, an extend gap penalty of 0.1, and a blosum similarity matrix to the *Streptomyces* thioesterase protein of GENPEPT record CAA71338 (version CAA71338.1 GI:2647975, hereby incorporated by reference in its entirety) was used. The alignment (Figure 4) revealed an alignment score of 955 and 145 conserved identities indicating significant similarity between the two compared sequences.

These analyses show that *dptD* and *dptH* encode thioesterase proteins, specifically, the proteins of SEQ ID NOS: 7-8.

EXAMPLE 4: Identification of a Daptomycin NRPS

A. *Identification of dptD as a daptomycin NRPS subunit*

The predicted translation products of the *dptD* DNA sequences described above (Examples 2 and 3) were inspected visually for the occurrence of various protein motifs described in the NRPS literature. A *dptD* condensation ("M") motif, indicative of a condensation domain, was identified at nucleotides 78488-78511 of SEQ ID NO: 1 (all of the nucleotide positions discussed in Examples 4-6 refer to SEQ ID NO: 1). See, e.g., Kleinkauf, H., et al., Eur. J. Biochem., 236, pp. 335-351 (1996) for the various motifs in the NRPS; and Pospiech, et al., Microbiol., 142, pp. 741-746 (1996). An ATP-binding ("C") motif was identified at nucleotides 79898-79930, an ATP-binding ("E") motif was identified at nucleotides 80453-80488, an ATPase ("F") motif was identified at nucleotides 80558-80581, and an ATP-binding ("G") motif was identified at nucleotides 0654-80677. These motifs collectively are indicative of an adenylation domain. A thiolation ("J") motif, indicative of a thiolation (PCP) domain, was identified at nucleotides 81050-81064. The above motifs, and the domains that they signify, belong to module 1 of *dptD*; in terms of Daptomycin synthetase, this is module 12.

Another *dptD* condensation (“M”) motif, indicative of a condensation domain, was identified at nucleotides 81623-81646. Another ATP-binding (“C”) motif was identified at nucleotides 83117-83149, an ATP-binding (“E”) motif was identified at nucleotides 83669-83704, an ATPase (“F”) motif was identified at nucleotides 83774-83797, and an ATP-binding (“G”) motif was identified at nucleotides 83870-83893. The above motifs collectively are indicative of another adenylation domain. Also a thiolation (“J”) motif, an indicator of a thiolation (PCP) domain, was identified at nucleotides 84257-84271. The above motifs, and the domains that they signify, belong to module 2 of *dptD*; in terms of Daptomycin synthetase, this is module 13.

The DptD amino acid sequences corresponding to the above-described predicted motifs and domains were identified (all of the amino acid positions for DptD refer to the amino acid positions in SEQ ID NO: 7). The motifs, and the domains that they signify, belonging to module 1 of DptD (corresponding to module 12 of Daptomycin synthetase) are as follows: A DptD condensation (“M”) motif was identified at coordinates 144-151; an ATP-binding (“C”) motif was identified at coordinates 614-624; an ATP-binding (“E”) motif was identified at coordinates 799-810; an ATPase (“F”) motif was identified at coordinates 834-841; an ATP-binding (“G”) motif was identified at coordinates 866-873; and a thiolation (“J”) motif was identified at coordinates 998-1002.

The DptD motifs, and the domains that they signify, belonging to module 2 of DptD (corresponding to module 13 of Daptomycin synthetase) are as follows: A DptD condensation (“M”) motif was identified at coordinates 1189-1196; an ATP-binding (“C”) motif was identified at coordinates 1687-1697; an ATP-binding (“E”) motif was identified at coordinates 1871-1882; an ATPase (“F”) motif was identified at coordinates 1906-1913; an ATP-binding (“G”) motif was identified at coordinates 1938-1945; and a thiolation (“J”) motif was identified at coordinates 2067-2071. The ATP-binding motifs are representative of adenylation domains.

B. *Identification of dptA, dptB and dptC as daptomycin NRPS subunits*

Certain M, C, E, F, G and J motifs were identified in a similar fashion in *dptA*, *dptB* and *dptC*. The sequence and type of each motif, the genes and modules in which each motif is found, as well as the amino acid and nucleotide coordinates of each motif, are shown below in Table 1:

Table 1

Gene	Module	Motif Type	Sequence	Amino Acid Coordinates	Nucleotide Coordinates
dptA	1	M	HHIALDGY	138-145	38966-38989
dptA	1	C	QTSGSTGRPKG	603-613	40361-40393
dptA	1	E	GELYLAGEGLAR	784-795	40904-40939
dptA	1	F	RMYRTGDL	819-826	41009-41032
dptA	1	G	RIELGEVQ	851-858	41105-41128
dptA	1	J	LGGHS	981-985	41495-41509
dptA	2	M	HHTAGDGA	1167-1174	42053-42076
dptA	2	C	YTSGSTGRPKG	1657-1667	43523-43555
dptA	2	E	GELHVAGEGLAR	1843-1854	44081-44116
dptA	2	F	RMYRTGDL	1878-1885	44186-44209
dptA	2	G	RIELGEVE	1910-1917	44282-44305
dptA	2	J	LGGDS	2041-2045	44675-44689
dptA	3	M	HHVILDGW	2751-2758	46805-46828
dptA	3	C	YTSGSTGLPKG	3238-3248	48266-48298
dptA	3	E	GELYVAGDGLAR	3420-3431	48812-48847
dptA	3	F	RMYRTGDL	3455-3462	48917-48940
dptA	3	G	RIELGEVE	3487-3494	49013-49036
dptA	3	J	LGGHS	3616-3620	49400-49414
dptA	4	M	HHIAGDGW	3806-3813	49970-49993
dptA	4	C	YTSGSTGRPKG	4292-4302	51428-51460
dptA	4	E	GEMYVAGAGLAR	4490-4501	52022-52057
dptA	4	F	RLYRTGDL	4525-4532	52127-52150
dptA	4	G	RIELGEIE	4557-4564	52223-52246
dptA	4	J	LGGHS	4688-4692	52616-52630
dptA	5	M	HHIAGDGW	4873-4880	53171-53194
dptA	5	C	HTSGSTGRPKG	5363-5373	54641-54673
dptA	5	E	GEIHIAGSGLAR	5553-5564	55211-55246
dptA	5	F	RMYRTGDL	5587-5594	55313-55336
dptA	5	G	RIELGDVE	5619-5626	55409-55432
dptA	5	J	LGGDS	5749-5753	55799-55813
dptB	1	M	HHVILDGW	142-149	56467-56490
dptB	1	C	HTSGSTGRPKG	611-621	57874-57906
dptB	1	E	GELYLAGTQLAR	803-814	58450-58485
dptB	1	F	RMYRTGDL	838-845	58555-58578
dptB	1	G	RIEPAEIE	870-877	58651-58674
dptB	1	J	AGGHS	998-1002	59035-59049
dptB	2	M	HHIAGDGW	1184-1191	59593-59616
dptB	2	C	YTSGSTGRPKG	1691-1701	61114-61146

dptB	2	E	GELYVAGVGLAR	1873-1884	61660-61695
dptB	2	F	RMYRTGDL	1908-1915	61765-61788
dptB	2	G	RVELGEVE	1940-1947	61861-61884
dptB	2	J	LGGHS	2069-2073	62248-62262
dptB	3	M	HHVAFDAM	2259-2266	62818-62841
dptB	3	C	YTSGSTGRPKG	2740-2750	64261-64293
dptB	3	E	GELYVAGVGLAR	2923-2934	64810-64845
dptB	3	F	RMYRTGDL	2958-2965	64915-64938
dptB	3	G	RVELGEVE	2990-2997	65011-65034
dptB	3	J	LGGDS	3118-3122	65395-65409
dptB	4	M	HHVVLDGW	3805-3812	67456-67479
dptC	1	C	YTSGSTGRPKG	178-188	68889-68921
dptC	1	E	GELYVAGVGLAR	360-371	69435-69470
dptC	1	F	RMYRTGDL	395-402	69540-69563
dptC	1	G	RVELGEVE	427-434	69636-69659
dptC	1	J	LGGHS	558-562	70029-70043
dptC	2	M	HHIAGDGW	748-755	70599-70622
dptC	2	C	YTSGSTGQPKG	1236-1246	72063-72095
dptC	2	E	GELYIAGDGLAR	1422-1433	72621-72656
dptC	2	F	RMYRTGDL	1457-1464	72726-72749
dptC	2	G	RVELGEVE	1489-1496	72822-72845
dptC	2	J	LGGHS	1618-1622	73208-73223
dptC	3	M	HHIAGDGW	1809-1816	73782-73805
dptC	3	C	YTSGSTGRPKG	2290-2300	75225-75257
dptC	3	E	GELYLAGAGLAR	2480-2491	75795-75830
dptC	3	F	RMYRTGDL	2515-2522	75900-75923
dptC	3	G	RVELGEVE	2547-2554	75996-76019
dptC	3	J	LGGDS	2677-2681	76386-76400

The amino acid coordinates refer to the amino acid sequence of each protein (DptA: SEQ ID NO: 9; DptB: SEQ ID NO: 11; DptC: SEQ ID NO: 13). The nucleotide position refers to the nucleotide position in SEQ ID NO: 1.

EXAMPLE 5: Amino acid pocket code annotation

The amino acid pocket code refers to a set of amino acid residues in the adenylation (A) domain that are believed to be involved in recognition and or binding of the cognate amino acid. The amino acid pocket code for the thirteen daptomycin synthetase modules are shown below (Table 2).

The amino acid pocket code for the daptomycin synthetase modules was identified by visual inspection of alignments created using MacVector 7.0 of the putative Dpt translation product aligned with NRPS A domains (amino acid binding pockets) as described in Stachelhaus, T., H. D. Mootz, and M. A. Marahiel (1999), The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases, Chemistry and Biology 6:493-505. See also Challis, G. L., J. Ravel, and C. A. Townsend (2000), Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains, Chemistry and Biology 7:211-224.

Table 2.

Protein	Module (Amino acid)	Pocket Code	Amino Acid Coordinates	Nucleotide Position
DptA	1 (Trp)	DVSSIGAV	649, 650, 653, 690, 711, 713, 734, 742	40499-40780
DptA	2 (Asn)	DLTKLGDV	1702, 1703, 1706, 1741, 1764, 1766, 1790, 1798	43658-43949
DptA	3 (Asp)	DLTKLGAV	3284, 3285, 3288, 3318, 3341, 3343, 3367, 3375	48404-48679

DptA	4 (Thr)	DFWSVGMV	4338, 4339, 4342, 4381, 4410, 4412, 4438, 4446	51566-51892
DptA	5 (Gly)	DILQLGVI	5409, 5410, 5413, 5452, 5479, 5481, 5503, 5511	54779-55087
DptB	1 (Orn)	DTWDMGYV	662, 663, 665, 704, 730, 732, 755, 763	58027-58332
DptB	2 (Asp)	DLTKLGAV	1737, 1738, 1741, 1771, 1794, 1796, 1820, 1828	61252-61527
DptB	3 (Ala)	DVVSAAFV	2786, 2787, 2790, 2824, 2847, 2849, 2873, 2881	64399-64686
DptB/ DptC	4(B)/1(C) (Asp)	DLTKLGAV	224, 225, 228, 258, 281, 283, 307, 315	69027-69302
DptC	2 (Gly)	DILQVGMI	1282, 1283, 1286, 1325, 1348, 1350, 1372, 1380	72201-72497
DptC	3 (Ser)	DVWHISLV	2336, 2337, 2340, 2379, 2404, 2406, 2429, 2437	75363-75668
DptD	1 (3-MG)	DLGKTGVI	659, 660, 663, 697, 720, 722, 746, 754	80033-80320
DptD	2 (Kyn)	DAWTTTGV	1733, 1734, 1737, 1775, 1796, 1798, 1820, 1828	83255-83542

The amino acid coordinates refer to the amino acid sequence of each protein (DptA: SEQ ID NO: 9; DptB: SEQ ID NO: 11; DptC: SEQ ID NO: 13; DptD: SEQ ID NO: 7). The nucleotide position refers to the nucleotide position in SEQ ID NO: 1.

Similarities between essentially the entire adenylation domains for aspartate and asparagine in the daptomycin gene cluster and for the adenylation domains for aspartate, asparagine and threonine in the CDA III NRPS of *Streptomyces coelicolor* are shown in Figure 10. Amino acids were aligned and the dendrogram was constructed using the MacVector. The nomenclature is as follows: the name of the gene--the module number in the gene--the amino acid activated (one letter code). The alignment shows that the adenylation domains for aspartate and asparagine in the daptomycin gene cluster are more similar to each other than they are to a domain from an unrelated amino acid such as threonine. Further, the alignment shows that the adenylation domains for aspartate and asparagine in the daptomycin gene cluster are more similar to each other than they are similar to the modules for aspartate and asparagine in Cda.

EXAMPLE 6: Identification of Epimerase Domains in Daptomycin NRPS

The amino acid sequences of DptA, DptB, DptC and DptD were inspected for sequences that are characteristic of epimerase domains. Epimerase domains are responsible for converting an L-amino acid to a D-amino acid and are typically encoded by approximately 1.4-1.6 kb of DNA.

It was expected that there would be a total of two epimerase domains in the daptomycin gene cluster, because it was known that daptomycin contained two D-amino acids, D-Ala and D-Ser. One epimerase domain was identified in each of module 8 (D-Ala) and module 11 (D-Ser). Module 8 and 11 are approximately 1.4 kb larger than modules that did not contain an epimerase domain (approximately 4.6 kb each for modules 8 and 11 compared to 3.2 kb each for modules not containing an epimerase domain). Further, modules 8 and 11 contain motifs that are indicative of an epimerase domain, including the motifs K, L, M, N, O, P and Q (see Kleinkauf and Von Dohren, 236: 335-351 (1996)). See Table 3.

Surprisingly, an epimerase domain was also identified in module 2. Module 2 is 1.6 kb larger than expected. Further, module 2 contains a number of motifs that are characteristic of an epimerase domain, including motifs K, L, M, N, O, P and Q. See Table 3. This unexpected finding suggests that the asparagine in daptomycin is in the D configuration.

Table 3

Gene	Mod	Motif Type	Sequence	Amino Acid Coordinates	Nucleotide Coordinates
dptA	2	K	RWPVVEWL	2100-2107	44852-44875
dptA	2	L	VRERHDAW	2146-2153	44990-45013
dptA	2	M	HHLVVDGVSWRIVLG	2237-2251	45263-45307
dptA	2	N	VVDVEGHGRN	2374-2383	45674-45703
dptA	2	O	TVGWFTSIYPVRL	2395-2407	45737-45775
dptA	2	P	PDQGLGY	2439-2445	45869-45689
dptA	2	Q	FGFNYLG	2467-2473	45953-45973
dptB	3	K	RWPVVEWL	3183-3190	65590-65613
dptB	3	L	VRDRHEAW	3229-3236	65728-65751
dptB	3	M	HHLVVDGVSWRVVLG	33315-3329	65986-66030

dptB	3	N	VVDVEGHGRN	3452-3461	66397-66426
dptB	3	O	TVGWFTSVYPVRV	3473-3485	66460-66498
dptB	3	P	PDQGLGY	3517-3523	66592-66612
dptB	3	Q	FGFNYLG	3545-3551	66676-66696
dptC	4	K	RWPVVEWL	2742-2749	76581-76604
dptC	4	L	VRDRHEAW	2788-2795	76719-76742
dptC	4	M	HHLVVDGVSWRVVLG	2874-2888	76977-77021
dptC	4	N	VVDVEGHGRN	3011-3020	77385-77417
dptC	4	O	TVGWFTSVYPVRV	3032-3044	77451-77489
dptC	4	P	PDQGLGY	3076-3082	77583-77603
dptC	4	Q	FGFNYLG	3104-3110	77667-77687

The amino acid coordinates refer to the amino acid sequence of each protein (DptA: SEQ ID NO: 9; DptB: SEQ ID NO: 11; DptC: SEQ ID NO: 13; DptD: SEQ ID NO: 7). The nucleotide position refers to the nucleotide position in SEQ ID NO: 1.

To confirm that the asparagine in daptomycin was in the D configuration, high pressure liquid chromatography (HPLC) was performed. A hexa-peptide containing the amino acids ornithine, glycine, threonine, aspartic acid, asparagine, and deacylated tryptophan (Trp-Asn-Asp-Orn-Gly-Thr) was isolated from daptomycin by degradation. The peptide above was analyzed by HPLC under conditions that would separate the peptide containing either the D-Asn or L-Asn. The HPLC showed only a single large peak for the isolated peptide above. See Figure 11, left panel. The peptide isolated from daptomycin was mixed with a peptide of the same sequence that had been synthesized in the laboratory and which contained D-Asn. The peptide mixture was analyzed by HPLC under the same conditions as before and shown to contain only a single peak. See Figure 11, middle panel. In addition, the peptide isolated from daptomycin was mixed with a synthetic peptide of the same sequence that contained L-Asn. HPLC analysis displayed two peaks. See Figure 11, right panel. These experiments confirm that naturally-occurring daptomycin contains D-Asn, not L-Asn.

From the experiments presented in Examples 2-7, the organization of the daptomycin NRPS was determined. Figure 12 shows the organization of *dptA*, *dptB*, *dptC* and *dptD*. *dptA* contains five modules (modules 1-5), *dptB* contains three modules (modules 6-8) and the catalytic domain of module 9, *dptC* contains the adenylation and thiolation domain of module 9 as well as two other modules (modules 10-11), and *dptD* contains two modules (modules 12-13) and a thioesterase domain.

Table 4 summarizes the correspondence between the 13 modules, their domains, the *dpt* genes, and their cognate amino acids. "C" represents a catalytic domain, "A" represents an adenylation domain, "T" represents a thiolation domain, "E" represents an epimerase domain, and "Te" represents a thioesterase domain.

Table 4.

Module	Cognate Amino Acid	Domains	Gene
01	L-Trp	CAT	*dptA*
02	D-Asn	CATE	*dptA*
03	L-Asp	CAT	*dptA*
04	L-Thr	CAT	*dptA*
05	Gly	CAT	*dptA*
06	L-Orn	CAT	*dptB*
07	L-Asp	CAT	*dptB*
08	D-Ala	CATE	*dptB*
09	L-Asp	CAT	*dptB/C*
10	Gly	CAT	*dptC*
11	D-Ser	CATE	*dptC*
12	L-MG	CAT	*dptD*
13	Kyn	CAT-Te	*dptD*

EXAMPLE 7: Transformation of Streptomyces lividans With The Daptomycin Gene Cluster From Streptomyces roseosporus

E. coli cells containing the BAC DNA from clone B12:03A05 (see Example 2) were grown in 5 mL of Luria Broth (LB; Difco) with agitation (250 rpm) overnight at 37°C. The BAC DNA was isolated by a standard alkaline lysis procedure (see Sambrook et al., *supra,* "Small scale preparation of plasmid DNA").

S. lividans TK64 spores were used to inoculate 25 mL of YEME + sucrose media and the culture was incubated for 40 hours at 30°C. The cultures were then harvested and the mycelium was pelleted away from the supernatant and washed several times with P-buffer (Practical *Streptomyces* Genetics; Tobias Kieser, Mervyn J. Bibb, Mark J. Buttner, Keith F. Chater and David Hopwood (John Innes Foundation, Norwich, 2000) ("the Hopwood Manual")). Fresh protoplasts were prepared according to the method described in the Hopwood manual (p. 56) and aliquoted into 0.5 mL portions (approximately 10^8-10^9 protoplasts) and pelleted by centrifugation at 3000 rpm for 7 minutes. Most of the supernatant was removed, leaving the pellet and approximately 50μL of the supernatant. The pellet was resuspended in the remaining supernatant, to which was added 5μL of BAC DNA from clone B12:03A05 (50 ng/μL in TE). This suspension was gently mixed before and after adding 350μL of a 25 % PEG-1000 in P-buffer solution (Hopwood Manual).

The protoplast suspension mixture was spread, in equal amounts, onto three dried R5T plates (dried to lose approximately 15% of their original weight; see Hopwood Manual). Inoculated plates were incubated overnight at 30°C. After 16-18 hours of growth, the plates were overlaid with 3 mL of an apramycin solution (1 mg/mL) in 20% glycerol to provide a final concentration of approximately 100μg/mL on each plate, and the plates incubated at 30°C. After three days, the plates were determined, by examination, to contain colonies which were growing in the presence of the apramycin selection. Two colonies were picked and streaked onto two F10A agar plates (2.5% agar, 0.3% calcium carbonate, 0.5% distillers solubles, 2.5% soluble starch, 0.5% yeast extract, 0.2% dextrose and 0.5% bactopeptone; suspended in 1 L deionized and autoclaved water) containing 100 μL/mL of apramycin and allowed to incubate at 30°C until the colonies sporulated. Spores were harvested according to the methods described in the Hopwood manual and stored as 20% glycerol suspensions at -20°C.

The spores derived from the transformation of *S. lividans* with BAC DNA containing the daptomycin gene cluster (from clone B12:03A05) were grown in an appropriate medium and analyzed by high pressure liquid chromatography (HPLC) and LC-MS to determine if they produced a wild-type lipopeptide profile (see Example 9).

EXAMPLE 8: Fermentation of Streptomyces lividans TK64 clone containing the daptomycin gene cluster

Spores of the *Streptomyces lividans* TK64 clone containing the daptomycin gene cluster (from clone B12:03A05) were harvested by suspending a 10 day old slant culture of medium A (2% irradiated oats (Quaker), 0.7% tryptone (Difco), 0.2% soya peptone (Sigma), 0.5% sodium chloride (BDH), 0.1% trace salts solution, 1.8% agar no. 2 (Lab M), 0.01 % apramycin (Sigma)) in 5 mL 10% aqueous glycerol (BDH)). 1 mL of this suspension, in a 1.5 mL cryovial, comprises the starting material, which was retrieved from storage at -135 °C. A pre-culture was produced by aseptically placing 0.3 mL of the starting material onto a slope of medium A1 and incubating for 9 days at 28 °C.

A seed culture was generated by aseptically treating the pre-culture with 4 mL of a 0.1 % Tween 80 (Sigma) solution and gently macerating the slope surface to generate a suspension of vegetative mycelium and spores. A two mL aliquot of this suspension was transferred into a 250 mL baffled flask containing 40 mL of nutrient solution S (1% D-glucose (BDH), 1.5% glycerol (BDH), 1.5% soya peptone (Sigma), 0.3% sodium chloride (BDH), 0.5% malt extract (Oxoid), 0.5% yeast extract (Lab M), 0.1 % Junlon PW100 (Honeywell and Stein Ltd), 0.1% Tween 80 (Sigma), 4.6% MOPS (Sigma) adjusted to pH 7.0 and autoclaved)) and shaken at 240 rpm for 44 hours at 30 °C.

Production cultures were generated by aseptically transferring 5% of the seed culture to baffled 250 mL flasks containing 50 mL medium P (1% glucose (BDH), 2% soluble starch (Sigma), 0.5% yeast extract (Difco), 0.5% casein (Sigma), 4.6% MOPS (Sigma) adjusted to pH 7 and autoclaved)) and shaken at 240 rpm for up to 7 days at 30 °C.

EXAMPLE 9: Purification and Analysis of the A21978C Lipopeptides from Fermentations of the Streptomyces lividans TK64 Clone Containing the Daptomycin Gene Cluster

Production cultures described in Example 8 were sampled for analysis by aseptically removing 2 mL of the whole culture and centrifuging for 10 minutes prior to analysis. Volumes up to 50 microlitres of the supernatant were analyzed to monitor for production of the native lipopeptides (A21978C) as produced by *Streptomyces roseosporus*. This analysis was performed at ambient temperature using a Waters Alliance 2690 HPLC system and a 996 PDA detector with a 4.6 x 50 mm Symmetry C8 3.5 µm column and a Phenomenex Security Guard C8 cartridge. The gradient initially holds at 90% water and 10% acetonitrile for 2.5 minutes, followed by a linear gradient over 6 minutes to 100% acetonitrile. The flow rate is 1.5 mL per minute and the gradient is buffered with 0.01% trifluoroacetic acid. By day 2 of the fermentation, production of three of the native lipopeptides, C1, C2 and C3, with UV/visible spectra identical to that of daptomycin, was evident, as shown by HPLC peaks with retention times of 5.62, 5.77 and 5.90 minutes (λmax 223.8, 261.5 and 364.5 nm) under the analytical conditions stated, as shown in Figure 5A. The lipopeptides then remained evident in the fermentation at each sample point during the 7-day period. Total yields of lipopeptides C1, C2 and C3 ranged from 10-20 mg per liter of fermentation material.

Liquid chromatography-mass spectrometry (LC-MS) analysis was performed on a Finnigan SSQ710c LC-MS system using electrospray ionization in positive ion mode, with a scan range of 200-2000 daltons and 2 second scans. Chromatographic separation was achieved on a Waters Symmetry C8 column (2.1x 50mm, 3.5μm particle size) eluted with a linear water-acetonitrile gradient containing 0.01% formic acid, increasing from 10% to 100% acetonitrile over a period of six minutes after a initial delay of 0.5 minutes, then remaining at 100% acetonitrile for a further 3.5 minutes before re-equilibration. The flow rate was 0.35 mL/minute and the method was run at ambient temperature.

The identification of the three native lipopeptides was confirmed, as indicated by molecular ions ($[M+H]^+$) at m/z of 1634.7, 1648.7 and 1662.7, which is in agreement with the masses reported for the major A21978C lipopeptide metabolites C1, C2 and C3, respectively, produced by *Streptomyces roseosporus* (Debono, M., et. al., J. Antibiotics, 40, pp. 761-777 (1987)).

Similar experiments were performed using the BAC clones 01G06, 06A12, 12F06 and 18H04. None of the *S. lividans* cells containing any one of these BAC clones were able to produce daptomycin.

EXAMPLE 10: Fed-batch fermentation of Streptomyces lividans TK64 Clone Containing the Daptomycin Gene Cluster for the production of Daptomycin

Cells of the *Streptomyces lividans* TK64 clone containing the daptomycin gene cluster (from clone B12:03A05) were regenerated by suspending a 10 day old slope culture of medium A (see Hopwood Manual; 2% irradiate oats (Quaker), 0.7% tryptone (Difco), 0.2% soya peptone (Sigma), 0.5% sodium chloride (BDH), 0.1% trace salts solution, 1.8% agar no. 2 (Lab M), 0.01% apramycin (Sigma) in 5 mL 10% aqueous glycerol (BDH)). A 1.5 mL cryovial containing 1 mL of starting material was retrieved from storage at -135 °C and thawed rapidly. A pre-culture was produced by aseptically placing 0.3 mL of the starting material onto a slope of medium A and incubating for 9 days at 28 °C. Material for inoculation of the seed culture was generated by aseptically treating the preculture with 4 mL of a 0.1 % Tween 80 (Sigma) solution and gently macerating the slope surface to generate a suspension of vegetative mycelium and spores.

A seed culture was produced by aseptically placing 1 mL of the inoculation material into a 2L baffled Erlenmeyer flask containing 250 mL of nutrient solution S (see Hopwood manual) shaken at 240 rpm for 2 days at 30 °C.

A production culture was generated by aseptically transferring the seed culture to a 20L fermenter containing 14 liters of nutrient solution P (see Hopwood manual). The production fermenter was stirred at 350 rpm, aerated at 0.5vvm, and temperature controlled at 30 °C. After 20 hours incubation a 50% (w/v) glucose solution was fed to the culture at 5 g/hr throughout the fermentation.

After 40 hours incubation, a 50:50 (w/w) blend of decanoic acid:methyl oleate (Sigma and Acros Organics, respectively) was fed to the fermenter at 0.5 g/hr for the remainder of fermentation. The culture was harvested after 112 hours, and the biomass removed from the culture supernatant by batch processing through a bowl centrifuge.

The biomass was discarded and the clarified fermentation broth was retained for extraction. The broth (approximately 10L) was loaded onto a 60 mm (diameter) by 300mm (length) column of HP20 resin, which had been pre-equilibrated with water, at a rate of 100 mL/min. The column was washed with 2L of water and then with 1.5L of 80% methanol (in water) at a similar flow rate. Finally, the bound material was eluted with 2L methanol and then taken to an aqueous concentrate under vacuum. The concentrate was diluted to 1L with purified water and partitioned with ethyl acetate (700 mL) three times. The ethyl acetate fraction was analyzed and discarded, and the aqueous layer was lyophilized to a powder. .

Daptomycin was isolated by high performance liquid chromatography (HPLC) using a radially compressed cartridge column consisting of two 40x100mm Waters Nova-Pak C18 6µm units and a 40x10mm Guard-Pak with identical packing. Lyophilized material (150 to 200mg) was dissolved in water and chromatographed on the columns using a gradient in which the initial conditions were 90% water and 10% acetonitrile, followed by a linear gradient over 10 minutes to 20% water and 80% acetonitrile, and then immediately ramping up to 100% acetonitrile over a further minute. UV absorption at 223nm was monitored for elution of daptomycin. The daptomycin peak eluted at about 9 minutes and was collected and combined over many repeated runs. The sample was then evaporated under vacuum and then dried *in vacuo* to yield 30 mg of purified compound. Only a proportion of the total material was processed.

The purified compound was first analyzed by reversed phase HPLC at ambient temperature on a 4.6 x 50 mm Waters Symmetry C8 3.5 µm particle size column with a Phenomenex Security Guard C8 cartridge using a Waters Alliance 2690 HPLC system and a 996 PDA detector. The column was eluted with a water-acetonitrile gradient, initially holding at 90% water for 2.5 minutes and then rising linearly over 6 minutes to 100% acetonitrile, at a flow rate of 1.5 mL/minute. The gradient was buffered with 0.01% trifluoroacetic acid. This chromatographic analysis confirmed that the retention time (5.52 mins) and the UV absorption spectrum (λ_{max} 223.8, 261.5, 366.9nm) of the purified compound matched those of daptomycin. LC-MS(ESI) confirmed the molecular ion MH^+ as 1620.6 (Figure 5B) and the ^{1}H NMR (D6-DMSO) gave a good visual match with that recorded for daptomycin (Figure 5C).

The identification of the material as daptomycin was further confirmed by ^{13}CNMR experiments, including DEPT and TOCSY.

Feed-batch fermentation may also be accomplished at a larger scale, for example at 60,000 liters.

EXAMPLE 11: The use of daptomycin genes for yield enhancement

A. Duplication of a positive regulatory gene

A neutral genomic site in the chromosome of *Streptomyces roseosporus* is identified by transposon mutagenesis with TN*5097*, or a related transposon, followed by fermentation analysis. The neutral site is excised from the chromosome using a restriction endonuclease that cuts outside of the neutral site and transposon, and cloned in *Escherichia coli*, selecting for the expression of the antibiotic resistance marker in the transposon (hygromycin resistance in the case of TN*5097*). An example of this approach was used to identify a neutral site in *Streptomyces fradiae*, the tylosin producer. See Baltz et al., Antonie van Leeuwenhoek, 71, pp. 179-187 (1997), incorporated herein by reference in its entirety. An example of identifying a neutral site in *S. roseosporus* is described in McHenney et al., J. Bacteriol., 180, pp. 143-151 (1998), incorporated herein by reference in its entirety.

The regulatory gene from the daptomycin gene cluster (SEQ ID NO:1) is cloned into a plasmid within the neutral site. A suitable plasmid would be one containing an antibiotic resistance gene for the selection of primary recombinants containing single crossovers, a counter-selectable marker such as the wild type *rpsL* gene, a ribosomal protein gene that confers sensitivity to streptomycin (Hosted and Baltz, J. Bacteriol., 179, pp. 180-186 (1997)) for selection of recombinants containing double crossovers that insert the cloned regulatory gene, and upstream and downstream sequences, into the chromosomal neutral site, and eliminate the plasmid sequences, and a thermal sensitive replicon that would facilitate the curing of the plasmid. The double crossover is done in a host strain that is normally resistant to streptomycin because it contains a mutation in the *rpsL* gene. Since the wild type (streptomycin-sensitive) allele of *rpsL* is dominant over streptomycin resistance, recombinants expressing streptomycin resistance must have eliminated the *rpsL* gene on the plasmid by a double crossover in the two arms of the neutral site, thus inserting the cloned daptomycin regulatory gene into the chromosome. Recombinants are fermented to verify that they produce an increased yield compared to the parental strain lacking the cloned daptomycin regulatory gene.

B. Duplication of ABC transporter genes

The pair of ABC transporter genes from the daptomycin gene cluster (SEQ ID NO:1), including upstream and downstream sequences, is cloned into the neutral site vector described above and inserted by double crossover into the *S. roseosporus* chromosome as described in Example 11A. Recombinants are fermented to verify that they produce increased levels of Daptomycin compared to the parental strain lacking the cloned ABC transporter genes.

C. Duplication of novA,B,C homologs

The segment of DNA containing the *novA,B,C* homology from the daptomycin gene cluster (SEQ ID NO:1), including the upstream and downstream sequences, is cloned into the neutral site vector and inserted by double crossover into the *S. roseosporus* chromosome as described in Example 11A. Recombinants are fermented to verify that they produce increased levels of Daptomycin compared to the parental strain lacking the cloned *novA,B,C* genes.

D. Duplication of daptomycin biosynthetic genes

The daptomycin biosynthetic genes, *dptA, B, C, D, E, F, G* and *H* (SEQ ID NO:1), including the fatty acyl-CoA ligase, the four subunits of the NRPS, the integral thioesterase of *dptD* and the free thioesterase of *dptH*, are cloned into a BAC vector that contains the fC31 attachment and integration functions (*att/int*) and oriT from plasmid RK2 (Baltz, Trends in Microbiology, 6, pp. 76-83 (1998), incorporated herein by reference in its entirety) for conjugation from *E. coli* to *S. roseosporus*. The BAC containing the daptomycin genes is introduced into *S. roseosporus* by conjugation from *E. coli* S17.1, or a strain containing a self-replicating plasmid RK2 (Id.). Alternatively, the BAC vector inserts into the chromosome by homologous recombination into the daptomycin gene cluster. Recombinants are fermented to verify that they produce increased levels of Daptomycin compared to the parental strain lacking the cloned daptomycin genes.

E. Duplication of daptomycin thioesterase genes

The daptomycin gene cluster (SEQ ID NO:1) contains at least two genes (*dptD* and *dptH*) having open reading frames (SEQ ID NO: 3 and SEQ ID NO: 6, respectively) or domains thereof that encode amino acid sequences which include conserved sequence motifs characteristic of proteins having thioesterase activity.. See SEQ ID NO:7 and SEQ ID NO:8 for DptD and DptH amino acid sequences, respectively. Either one (or both) of these thioesterase genes or the thioesterase domains thereof can be duplicated by following the procedure of Example 11A, above.

A segment of DNA containing the *dptD* ORF sequences (e.g., SEQ ID NO: 1; SEQ ID NO:3) optionally linked in an operative fashion to an expression control sequence (such as the natural one in SEQ ID NO:1 or 2) and optionally including the upstream and downstream sequences, is cloned into a neutral site vector and inserted by double crossover into the *S. roseosporus* chromosome as described in Example 11A. Recombinants are fermented to verify that they produce increased levels of Daptomycin compared to the parental strain lacking the cloned *dptD* gene.

Similarly, a segment of DNA containing the *dptH* ORF sequences (e.g., SEQ ID NO:4, SEQ ID NO:6) optionally linked in an operative fashion to an expression control sequence (such as the natural one in SEQ ID NOS:1, 4 or 5) and optionally including the upstream and downstream sequences, is cloned into a neutral site vector and inserted by double crossover into the *S. roseosporus* chromosome as described in Example 11A. Recombinants are fermented to verify that they produce increased levels of Daptomycin compared to the parental strain lacking the cloned *dptH* gene.

Other suitable hosts (i.e., those having NRPS or PKS multienzyme complexes) may be transformed with segments of DNA encoding proteins from the daptomycin gene cluster having thioesterase activity for improved peptide production. Alternatively, polypeptides encoded by such segments of DNA may be introduced into *S. roseosporus* or said other suitable hosts by protein transfer techniques well-known to those of skill in the art.

F. Duplication of daptomycin resistance genes

The daptomycin resistance gene(s) are identified by cloning and expression in an appropriate streptomycete host that is naturally susceptible to Daptomycin. The cloned daptomycin resistance gene(s) are inserted into the neutral site vector within the neutral site, and inserted into the *S. roseosporus* chromosome by double crossover as described in Example 11A. Recombinants are fermented to verify that they produce increased levels of Daptomycin compared to the parental strain lacking the cloned daptomycin resistance genes.

G. Duplication of daptomycin biosynthetic genes and accessory genes

The complete set of daptomycin biosynthetic genes such as those contained on the BAC clone B12:03A05 (see Example 2 and SEQ ID NO:1) are introduced into *S. roseosporus* by conjugation from *E. coli* (or by another method of DNA-mediated transformation) and inserted into the chromosome by site-specific integration into the ϕC31 integration site as in Example 11D, leading to a duplicate version of the daptomycin biosynthetic and accessory genes. Alternatively, the BAC vector inserts into the chromosome by homologous recombination into the daptomycin gene cluster (as verified, e.g., by Southern blot analyses), leading to tandem duplication of the daptomycin biosynthetic and accessory genes at their native location. Recombinants are fermented to verify that they produce increased levels of daptomycin compared to the parental strain lacking the cloned daptomycin genes and accessory genes.

EXAMPLE 12: The Use of Daptomycin Biosynthetic Genes To Produce Novel Products

A. Modification of the peptide structure by site-directed mutagenesis of an amino acid specificity code: conversion of position 2 D-Asn to D-Asp.

The amino acid specificity codes for the thirteen amino acids in Daptomycin are shown in Table 1 (see Example 6, above). See also Stachelhaus et al., Chem. Biol., 6, pp. 493-505 (1999), incorporated herein by reference in its entirety, for a discussion of identifying and altering adenylation domain amino acid specificity codes in NRPSs. The code for all three L-asp residues in positions 3, 7, and 9 of daptomycin are identical: DLTKLGAV (where the letters indicate standard amino acid abbreviations). The code for D-Asn in position 2 is DLTKLGDV, and it differs by a single amino acid (a D instead of A in position 7). The D-Asn specificity code is changed to that specifying D-Asp by making a site specific change in the adenylation domain of module 2 in PS I.

The mutant version of module 2 is inserted into the *S. roseosporus* chromosome by gene replacement (see Example 11). A counter selectable marker (*e.g.*, the wild type *rpsL* gene) is inserted into the adenylation domain of module 2 by gene replacement. The mutant module 2 adenylation domain containing the coding sequence for D-Asp, and containing flanking DNA (about 1 to 5 kb on each side of the specificity code) on an appropriate thermal sensitive plasmid is introduced into the *S. roseosporus* strain disrupted for daptomycin biosynthesis. Recombinants containing single crossovers are selected at the non-permissive temperature by selection for an antibiotic resistance marker on the plasmid (*e.g.*, hygromycin, apramycin or thiostrepton resistance). If the host strain is streptomycin resistant by a mutation in the chromosomal *rpsL* gene, then the second crossover completing the gene replacement can be selected for streptomycin resistance. The recombinant is screened for antibiotic production. The novel derivative of Daptomycin is separated and analyzed to confirm the structure according to methods described, e.g., in United States Patents RE 32,333, RE 32,455, 4,874,843, 4,482,487, 4,537,717, and 5,912,226.

B. Molecular exchange of an amino acid coding module for one of different amino acid specificity.

Daptomycin has four acidic amino acids: three L-asp residues at positions 3, 7, and 9, and a 3-methyl-Glu (3-MG) at position 12 (see Table 1, Example 6). Novel derivatives of Daptomycin are generated by exchanging the adenylation domain that specifies 3-MG for one that specifies L-asp. The adenylation domain of the 3-MG module is inserted into segments of the L-asp module flanking the L-asp adenylation domain which has been removed by molecular genetic procedures. The hybrid 3-MG module containing the flanking DNA from an L-asp module is inserted into an appropriately constructed gene replacement vector, and the hybrid module is exchanged for an L-asp module by homologous double crossover as in Example 11A. This same procedure is repeated for the other two L-asp modules. The recombinants produce three novel derivatives of Daptomycin containing 3-MG substituted for L-asp in positions 3, 7, or 9, and maintain the overall four negative charges in the molecule.

C. Exchange of a non-ribosomal peptide synthetase (NRPS) subunit for one that catalyzes the incorporation of different amino acid(s).

The gene that encodes the fourth subunit of the Daptomycin NRPS (PS-IV; see Table 1, Example 6) contains two modules that encode the specificity for incorporation of amino acids 12 (3-MG) and 13 (L-kyn). The gene that encodes the third subunit for the biosynthesis of the cyclic lipopeptide CDA (Kempter et al., Angew. Chem. Int. Ed. Engl. , 36, pp. 498-501 (1997); Chong et al., Microbiology, 144, pp. 193-199 (1998); each of which is incorporated by reference herein in its entirety) also encodes the last two amino acids, in this case amino acids 10 (3-MG) and 11(L-trp). A derivative of Daptomycin containing L-trp instead of L-kyn in position 13 is generated by disrupting gene *dptD*, and by replacing it with the gene that encodes PSIII for CDA. Expression of the PSIII gene from a strong promoter (*e.g.*, the ermEp* promoter; Baltz, Trends in Microbiology, 6, pp. 76-83 (1998), incorporated herein by reference in its entirety), and inserted into a neutral site in the *S. roseosporus* genome as described in Example 11A, allows CDAPSIII to complement the *dptD* mutation and results in the production of the altered daptomycin with L-trp replacing L-kyn. The recombinant is fermented and the product(s) of the recombinant are analyzed by LC-MS as described in Example 9.

D. Insertion of an extra internal module to cause the expansion of the Daptomycin ring from 10 amino acids to 11 amino acids or lengthening of the tail to 4 amino acids.

A simple NRPS elongation module may be defined as comprising domains "C-A-T" (condensation-, adenylation- and thiolation-domains). To link modules, and to identify a permissive site within the Daptomycin NRPS in which to insert additional internal modules, the domain and inter-domain regions are examined for sequences indicative of flexible "linker" sequences. See, e.g., Mootz et al., Proc. Natl. Acad. Sci.

U.S.A., 97, pp. 5848-5853 (2000), which is incorporated herein by reference in its entirety. Sequences encoding an additional module are inserted in the linker sequence between an upstream T-domain and a downstream C-domain using well-known genetic recombination techniques, e.g., see Example 11A, above.

Isolation of the module DNA is obtained from the chromosomal DNA extracted from the producer organism. Various isolation techniques can be used such as, cutting the chromosomal DNA with restriction enzymes and isolating a fragment coding for the module of interest after it is identified by means of Southern blot or isolation of the module of interest by genetic amplification (PCR) using suitable primers. Sequencing and characterization of the amplified fragments as well as cloning can be performed according to conventional techniques. New modules can be inserted between the modules specifying L-Thr and Gly in *dptA*; between the modules specifying L-Orn and L-Asp or L-Asp and D-Ala in *dptB*; between L-Asp and Gly or Gly and D-Ser in *dptC*; and between modules specifying 3-MG and L-Kyn in *dptD* to expand the ring of daptomycin. New modules can be inserted in the *dptA* gene between the modules specifying L-Trp and D-Asn, D-Asn and L-Asp, or L-Asp and L-Tyr to lengthen the tail of daptomycin. The module insertions can be carried out using the methods for double crossovers described in Example 11A.

E. Insertion of an additional carboxyl terminus module adjacent to and upstream from the thioesterase module.

Carboxy-terminal thioesterase domains ("Te-domains") of a variety of NRPSs and PKSs can cleave (i.e., catalyze chain termination) non-natural peptide and polyketide substrates. See Mootz et al., *supra*; see also de Ferra et al., J. Biol. Chem., 272, 25304-25309 (1997); each of which is hereby incorporated by reference in its entirety. Te-domains can act as hydrolases, releasing a linear product, or as cyclases, releasing cyclic peptides. Evidence suggests that a Te-domain which functions as a cyclase in its natural configuration within a NRPS or PKS may, nonetheless, function as a hydrolase when engineered into new modular configurations. (An isolated C-terminal Te-domain has been shown to catalyze cyclization on various substrates as long as key "recognition amino acids" are at the C- and N-termini of the substrate; see Trauger et al., Nature, 407, pp. 215-218 (2000).)

It has also been shown that some C-terminal Te-domains function best, when moved, by retaining their association with a portion of the protein domain occurring directly upstream in the natural NRPS or PKS modular configuration. See Guenzi et al., J. Biol. Chem., 273, pp. 14403-14410 (1998), incorporated herein by reference in its entirety. It is possible that retaining the boundary between the Te-domain and a portion of the domain directly upstream (N-terminal) may also contribute to retaining cyclase function of the Te-domain within a new modular configuration.

Accordingly, to insert an additional module upstream from a Te-domain and have it be operatively linked thereto, one can identify linker sequences between the C-A-T modules and the C-terminal Te-domain, as described above, and insert sequences encoding the additional module therein, using standard genetic manipulations. Optionally, one can engineer a new, hybrid C-terminal Te-domain in which the C-terminal portion of the penultimate thiolation (T-) domain remains linked (or is otherwise grafted) to the Te-domain ("T-/Te- domain"). See Guenzi et al., 1998, *supra*. Sequences encoding the additional module are then inserted within the identified linker region upstream from a hybrid T-/Te domain using well-known genetic recombination techniques, as described in Example 11A, above

F. Internal deletion of a module to cause the contraction of the Daptomycin ring from 10 amino acids to 9 amino acids or shortening of the tail.

To obtain a deletion of an internal module on the chromosome by double crossing-over and selection on antibiotic plates it is necessary to prepare a plasmid containing a fragment of chromosomal DNA situated upstream from the module to be deleted fused by ligation to a fragment downstream of the module to be deleted. The plasmid also carries a wild type *rpsL* gene to confer streptomycin sensitivity on recombinants in a streptomycin-resistant genetic background (see Example 11A), an antibiotic resistance gene (e.g., apramycin resistance, thiostrepton resistance or hygromyicin resistance) for selection of single crossovers, and a temperature sensitive replicon that can be cured at elevated temperature. A single crossover inserting the plasmid by homologous recombination into the region of DNA upstream of the module to be exchanged can be selected for antibiotic resistance at elevated temperature. The second crossover that deletes the module can then be selected on media containing streptomycin (thus eliminating all plasmid sequences). Recombinants containing deletions of the appropriate module can be verified by Southern blot hybridization of *S. roseosporus* DNA cleaved with appropriate restriction endonucleases. This approach can be taken to delete the L-Asp module from *dptB* or the Gly module from *dptC*, for example. It can also be used to delete the modules in the *dptA* gene specifying L-Asn, L-Asp or both L-Asn and L-Asp together.

G. Translocation of the terminal thioesterase module to cause the contraction of the Daptomycin ring.

Sequences encoding the thioesterase (Te) region which resides at the carboxyl terminus of the last module in the daptomycin NRPS (DptD) may be translocated upstream to the end of an internal module encoding region. This translocation will result in the release of a defined shortened product that will yield a truncated linear or cyclic peptide. The translocation of the Te can be accomplished by double crossovers much the same way as described above in Examples 12A and 12F.

H. Molecular exchange between Daptomycin NRPS and other NRPS or PKS genes

a. Dap thioesterase onto different NRPS or PKS

Using well-known molecular and genetic methods such as those described above, sequences encoding a C-terminal Te-domain of the daptomycin NRPS of the invention (e.g., DptD) may be moved (either alone or in combination with one or more upstream modules or portions thereof) into association with sequences encoding other NRPS or PKS modular genes from a variety of other hosts to produce hybrid modular synthetases that are capable of producing new peptide and/or hybrid peptide/polyketide products having useful properties. See, e.g., Stachelhaus et al., Science, 269, pp. 69-72 (1995) and Cane and Khosla, Chem. Biol., 6, pp. 319-325 (1999); each of which is incorporated herein by reference in its entirety. Similarly, daptomycin sequences encoding a free thioesterase of the invention (e.g., DptH) may be moved into association other NRPS or PKS encoding modular genes to produce hybrid modular synthetases.

b. Module and domain exchanges between dap and other NRPS and/or PKS

Various sequences derived from the daptomycin biosynthetic gene cluster of the invention -- including but not limited to domains and modular structures -- may be used to construct plasmids and other vectors for use in genetic recombination reactions (gene duplication, conversion, replacement, etc.) between daptomycin sequences and natural or synthetic NRPS and PKS sequences in homologous and heterologous hosts to produce hybrid NRPS and hybrid NRPS/PKS modular synthetases comprising sequences from the daptomycin biosynthetic gene cluster. Such hybrid synthetases will produce novel peptide and polyketide products which are expected to have new and useful properties.

I. Creation of Lipopeptide Derivatives of Nonribosomally-synthesized Peptides That Are Not Normally Acylated.

The fatty acid tail of daptomycin is thought to be attached by the products of the *dptE* and *dptF* genes, working in conjunction with the condensation domain at the start of *dptA*. These genes and gene fragments may be transferred to the beginning of a foreign nonribosomal peptide synthase gene, or to an internal location within the daptomycin gene cluster, either at the start of a gene (e.g. 5' of *dptB*, *C*, or *D*) or within a gene at the start of a module (e.g. 5' of module 2), to create acylated versions of the foreign nonribosomally synthesized peptide, or to create acylated, truncated derivatives of daptomycin. The foreign gene may be derived from another natural organism, or one generated by recombinant techniques, e.g. various versions of daptomycin that have undergone modifications to expand or contract the ring, to have substituted amino acids in the peptide sequence as described herein.

J. Modification of amino acid stereoisomers in the peptide structure.

Stereospecificity in the amino acid backbone produced by an NRPS is determined by the presence of epimerase domains in the donor module and distinctive condensation domains in the acceptor module. An alteration in stereochemistry of the amino acids may be achieved by addition of an epimerase domain to a donor module, and substitution of the appropriate condensation domain to the acceptor module. An alteration can also be made by removal of the epimerase domain from a donor module, and the substitution of the appropriate condensation domain in the acceptor, e.g. the epimerase domain can be excised from module 2 of *dptD*, and the condensation domain of module 3 of *dptD* can be exchanged for the condensation domain from another module that does not normally accept a D-amino acid. Useful epimerase and condensation domains may be found in the daptomycin cluster as well as in other nonribosomal peptide synthetase genes.

EXAMPLE 13: Procedure for Making a Linear Thioester That Can Be Cyclized to Daptomycin

A. Synthesis of pantetheine derivative of the Daptomycin linear peptide.

Pantetheine is obtained by the method of Overman (Overman, et al., 59 (1974)) from commercially available pantetheine. A column is loaded with a 2-chlorotrityl resin. Protected kynurenine (α-amino protected with 9-Fluorenylmethoxycarbonyl (Fmoc) aromatic amine protected with t-Boc) and its protected Cs salt are prepared and dissolved in N,N-Dimethyl formamide (DMF). This solution is added to a suitably prepared 2-chlorotrityl resin. The reaction proceeds until the protected kynurenine has been loaded onto the resin. The resin is washed to remove any unused reagent and CsCl salt.

Following is the iterative addition of the other 12 amino acids. This is the sort of process that may be done on an automated flow through system. The non α-carboxylic acids are protected as their trityl ester, hydroxyl groups are protected as acetyl esters, the other than α-amines are protected by t-Boc groups. α-amino groups are protected with Fmoc groups, except for acylated tryptophan, which is protected by the acyl group. A 0.02 M tetra-n-butylammonium fluoride trihydrate in DMF is added to cleave the Fmoc group of the resin bound growing peptide. The progress of the reaction is monitored through uv/vis absorption changes due to released Fmoc groups. The resin is rinsed to remove excess reagent.

To couple the next amino acid, the next suitably protected amino acid is dissolved in DMF to get a 0.1 M solution in DMF with 1 eq of Diisopropylcarbodiimide (DIPCDI) and 1 eq N-Hydroxybenzotriazole (HOBt). The reaction is allowed to proceed to completion. The resin is washed with DMF to insure that any excess reagents are removed.

This process is repeated until the peptide L-Kynurenine (t-Boc protected amine)-L-threo-3-methyl Glu (trityl ester)-D-Ser(acetyl ester)-Gly-L-Asp(trityl ester)-D-Ala-L-Asp(trityl ester)-L-Asp(trityl ester)-L-Orn(t-Boc protected)-Gly-L-Thr(acetyl ester)-L-Asp(trityl ester)-D-Asn-L-acylated tryptophan is obtained.

To obtain cleavage of the protected peptide, a 1:1:3 solution of acetic acid:trifluorethanol; Dichloromethane (DCM) is added to the resin and allowed to stand for 3 hours at 24°C. The protected peptide is precipitated with hexane and the solvent removed in vacuo. The solid is dissolved in tetrahydrofuran (THF) or other appropriate solvent. A 1.2 eq of Dicyclohexylcarbodiimide (DCC) and 1.2 eq of HOBt 1.2 eq of p-nitrophenol is added. After the reaction is completed, 2.5 eq of the sodium salt of pantetheine is added and stirred for as long as necessary for the reaction to go to completion. The crude reaction is chromatographed to yield the protected pantetheine thioester. The protected peptide is dissolved in a 16:3:1 solution of trifluoroacetic acid: DCM: pantetheine and allowed to stir for 3 hours at 24° C. It is precipitated with diethyl ether, dried and purified by preparative HPLC.

EXAMPLE 14: Using the Daptomycin Thioesterase to Build a Synthesis Based Drug Discovery Program (With Ultra-High-Throughput Screening Method)

A. Conversion of a lipopeptide synthesis program into a drug discovery program.

Photocleavable resins are available commercially and can be used in the preparation of a library of linear thioester containing peptides that are tethered to the resin by a photocleavable linkage. These linear thioesters are cyclized on resin to yield cyclic lipopeptides that could be cleaved by photolysis to yield lipopeptides of distinct molecular weight. The molecular weight of each member of the library is determined. These resin beads are encapsulated in an alginate matrix (macrodroplet) with a tester strain and a live or dead strain or some other colorimetric or fluorometric indicator of viability. After an empirically determined growth period the resin is illuminated at 365 nm to release the lipopeptide into the macrodroplet. If a given lipopeptide has bactericidal biological activity, then the cells die, leaving the macrodroplet colorless. Since the resin bead is spherical and the illumination source is unidirectional, there is approximately half of the lipopeptide material left on the resin bead. The alginate matrix is dissolved, the bead washed and agitated under illumination to yield the active molecule, whose identity is determined by LC-MS. By this method, a large library of synthetic compounds is screened rapidly and efficiently.

There will be some constraints on how the peptide is linked to the resin, for the thioesterase has to be able to cyclize it. This can be accomplished by using the lipid tail as a resin attachment site. By using a long chained carboxylic acid such as sebacic acid ($HO_2C(CH_2)_8CO_2H$), one side of the carboxylic acid is attached to the photocleavable resin via the amino group of an o-nitrobenzylamine, leaving the other free to build the peptide. This leaves enough freedom to allow for cyclization. An alternative method is to use a resin that has a long alkyl or polyether attachment site, which allows the peptide to be cyclized without interference from the bulky resin. The attachment site is varied so that a future asparagine or glutamine is attached to the orth-o-nitrobenzylamine of the photocleavable resin. Upon photocleavage the corresponding asparagine or glutamine is liberated. This would allow the cyclization to occur on the resin.

EXAMPLE 15: Using an Appropriate Synthetic Molecule To Isolate A Presumed, But Uncharacterized Thioesterase

A plasmid, suitable for library construction, expressible in *E. coli*, that secretes a cloned peptide into the medium is used. A desirable but uncharacterized thioesterase is selected and a DNA library is prepared from either the entire organism or a subset of the entire organism in the described plasmid. A suitable resin-bound linear thioester peptide is prepared that upon cyclization and cleavage yields the desired cyclic lipopeptide. The *E. coli* would have to be resistant to the cyclization product. The *E. coli* library is encapsulated in an alginate matrix along with one or more resin beads, such that only one *E. coli* clone was in each macrodroplet. The *E. coli* is grown for an empirically determined period in a pre-determined medium, so that sufficient secreted enzyme is present to cyclize the resin bound compound. The macrodroplets are placed on an appropriate target lawn and illuminated with 365 nm light. Those macrodroplets containing *E. coli* producing a secreted active thioesterase are readily identified by clearing zones surrounding the macrodroplet. The alginate macrodroplet is dissolved to yield the desired *E. coli* clone, which are then isolated and further evaluated. See, Trauger J. W., *et al, Nature*, 407: 215-18, 2000).

EXAMPLE 16: Use of free thioesterase

A. Expression of dptD or dptH related sequences in homologous or heterologous systems to increase efficiency of product formation by modular NRPSs and PKSs

The C-terminal Te-domain excised from tyrocidine synthetase has been shown to catalyze cyclization on various peptide substrates as long as key “recognition amino acids” are at the C- and N-termini of the substrate. See Trauger et al., Nature, 407, pp. 215-218 (2000), incorporated herein by reference in its entirety. Sequences derived from the C-terminal domain of daptomycin NRPS (e.g., *dptD*) may similarly be isolated and expressed – alone or in the form of suitable fusion proteins – in a homologous or heterologous host (or *in vitro* system) to catalyze cyclization of peptide and polyketide products which naturally (or which have been engineered to) possess key substrate recognition amino acids required for the daptomycin Te-domain to bind and join substrate ends (see below).

As discussed *supra* (Example 13), when isolating sequences derived from the C-terminal Te-domain of daptomycin synthetase (NRPS) for independent expression, it may be preferable to include natural C-terminal sequences from the penultimate amino acid module. See, e.g., Guenzi et al., 1998, *supra*. Various *dptD* and upstream-derived sequence combinations can be tested using techniques well-known in the art to optimize the thioesterase activity of the C-terminal Te-domain of daptomycin NRPS when expressed independently from upstream polypeptides such as DptA, DptB and/or DptC. Independent expression of the C-terminal Te-domain of daptomycin may be accomplished using standard molecular biology techniques. Independent expression of the C-terminal Te-domain of daptomycin NRPS is accomplished by inserting sequences derived from the thioesterase domain of the *dptD* ORF (SEQ ID NO:3) downstream from natural daptomycin NRPS promoter sequences (SEQ ID NO:2) in an appropriately constructed expression vector. Alternatively, independent expression of the C-terminal Te-domain of daptomycin NRPS is accomplished by inserting the thioesterase domain of the *dptD* ORF (SEQ ID NO:3) downstream from a heterologous promoter, which is constitutively active or from a heterologous promoter which may be turned on or off in a regulated manner. Those of skill in the art will appreciate the factors to be considered in selecting appropriate promoters and vectors for expression or over-expression in a host-dependent manner.

Sequences derived from the free thioesterase domain of the daptomycin biosynthetic gene cluster of the invention (*dptH*) may be similarly expressed in a homologous or heterologous host to test and develop novel cyclic peptides and the like.

The key recognition amino acids of daptomycin are identified by systematic mutagenesis of the amino acid residues of daptomycin followed by cyclization assays using each modified daptomycin substrate in a reaction catalyzed by the isolated Te-domain. C- and N-terminal amino acid residues required for daptomycin cyclization are identified and engineered into new substrate backbones into which peptide and polyketide building block units can be inserted. Substrate engineering can be performed at the nucleic acid sequence level or at the peptide level using techniques well-known to those of skill in the art. The length and composition of preferred substrates may be determined empirically, taking into consideration factors well-known to the skilled worker and including (but not limited to) substrate binding efficiency, catalytic rate, biological activity of resulting cyclic product(s), and ease of purification of the final products.

B. Mutagenize dptD or dptH to affect proof-reading function

The *dptH* gene from the daptomycin gene cluster is related to free thioesterase enzymes which are known to participate in the biosynthesis of some peptide and polyketide secondary metabolites. See e.g., Schneider and Marahiel, Arch. Microbiol., 169, pp. 404-410 (1998), and Butler et al., Chem.& Biol., 6, pp. 87-292 (1999), hereby incorporated by reference in their entirety. It has been suggested that editing thioesterases are often required for efficient natural product synthesis. Butler et al. have postulated that the free thioesterase found in the polyketide tylosin gene cluster may be involved in editing and proofreading functions, consistent with the suggested role of the thioesterases in efficient product formation.

As described in Example 13A, homologous or heterologous expression of the daptomycin *dptH* (encoding a free thioesterase) or the thioesterase-encoding domain of *dptD* (encoding the C-terminal Te) genes may affect the efficiency of product formation by modular NRPSs and PKSs. The proposed editing and proofreading functions of the daptomycin thioesterase type II enzyme (DptH) (and potentially of the type I thioesterase enzyme when detached from the C-terminus of the daptomycin gene cluster and separately expressed) may be altered by conventional mutagenesis and other recombinant DNA techniques, e.g., those known to affect adversely the fidelity of DNA replication. Altered and mutated forms of thioesterase genes may be expressed in appropriate expression systems and screened for those which encode thioesterase enzymes having altered biological properties. Especially desirable would be thioesterase enzymes that have higher than normal rates of amino acid misincorporation. Such mutants would be useful for creating a larger diversity of peptide and peptide/polyketide hybrid products having new and useful biological properties.

EXAMPLE 17: Using an Appropriate Synthetic Molecule To Test NRPS Thioesterase Activity Of Fragments, Muteins, Derivatives, Analogs And Homologous Proteins

A thioesterase fusion polypeptide, fragment, mutein, derivative, analog or homologous protein having potential thioesterase activity associated with a NRPS may be compared to a corresponding wild-type thioesterase polypeptide (e.g., from which it was derived) by transforming a suitable heterologous host cell independently with expression plasmids having nucleic acid sequences encoding the wild-type and the potential thioesterase polypeptides. Culturing the transformed host cells allows expression of the nucleic acid sequences, and the products of the NRPS may be purified and analyzed according to procedures well known to those of skill in the art. (Alternatively, homologous host cells in which one or more genes necessary for NRPS activity have been disabled or deleted may be used). The methods set forth in Examples 7-9 for analyzing daptomycin lipopeptide production in a heterologous host may be used in modified forms, for example, to monitor peptide production from a modified daptomycin or other NRPS comprising a thioesterase fusion, fragment, mutein, derivative, analog or homolog. Other cell growth or viability-based inhibition assays, such as that described in Example 15 for *E. coli*, may be used to monitor antibiotic, antifungal, antiviral, anticancer or other anti-cellular growth activities of peptides secreted by one host that may affect cell division, growth or viability of a second cell. Such secretion assays may be appropriately designed and modified to test the ability of a thioesterase to release from a NRPS a linear or cyclic peptide having anti-cellular growth activity. Once designed and optimized for sensitivity, such a secretion assay may then be used to compare systematically the ability of altered or mutated forms of a thioesterase to support the release of the same peptide from the NRPS.

EXAMPLE 18: Using Daptomycin Biosynthetic Genes to Identify and Isolate Related Genes

The nucleic acid and amino acid sequences of the invention can be compared to the corresponding sequences from another lipopeptide pathway in order to identify features that can then be used to identify sequences from an NRPS or a component of an NRPS encoding another lipopeptide.

The amino acid 3-methyl glutamic acid (3MG) is uncommon, but is found in daptomycin, the calcium dependent antibiotic (cda) from *S. coelicolor*, and the A54145 compound made by *S. fradiae*. Comparison of the *S. roseosporus* and *S. coelicolor* nucleic acid sequences that encode the 3MG adenylation domain, as well as from analogous sequences from genes that adenylate other amino acids, were used to create the primer pair P140 and P141:

P140 ACSSWSGGSGTSSCCTTCATGAA

P141 ATGGTGTTCGAGAACTAYCC.

An *S. fradiae* cosmid library was screened by PCR using P140 and P141 using standard techniques. The PCR reaction yielded a nucleic acid molecule product of approximately 700 bp, whose sequence proved similar to the region encoding the 3MG adenylation domain in *S. roseosporus* and *S. coelicolor*. Extension of the sequence by primer walking confirmed that the region identified was the 3MG module in A54145.

This method was also used to identify portions of an NRPS pathway that encode condensation domains downstream of a D-amino acid activating module. D-amino acids are unusual amino acids found in non-ribosomally synthesized peptides, and primers for condensation domains associated with them can be used to identify pathways with such amino acids. The nucleic acid sequences of the *S. roseosporus* daptomycin and *S. coelicolor* cda sequences that encode these D-amino acid condensation domains were compared to each other and to analogous sequences from other condensation domains associated with L-amino acids in order to create the primer pair P144 and P145:

P144 SCSCTSCAGGAGGGSHTSSTSTTCC

P145 CCGAASACSACGTCGTCSCGSCC.

An *S. fradiae* cosmid library was screened by PCR using P144 and P145 using standard techniques. The PCR reaction yielded a nucleic acid molecule products of approximately 800 basepairs, the sequences of which proved to be similar to the condensation domains following the D-amino acids in *S. roseosporus* and *S. coelicolor*. Sequences corresponding to more than one domain were obtained, indicating that the pathway had more than one D-amino acid.

These approaches, based on understanding the sequence of the daptomycin pathway, can be used to develop special primer sets for other genetic features of lipopeptide pathway gene clusters, such as regions encoding epimerase domains or the condensation domain of the first adenylation module responsible for condensing the fatty acid to the peptide, as well as genes involved in acylation, such as DptE and F.

Table 5

ORF# - Fragment	Nucleotide Sequence SEQ ID NO:	Amino Acid Sequence SEQ ID NO:
1 - 90 kb*	20	19
2 - 90 kb	22	21
3 - 90 kb	24	23
4 - 90 kb	26	25
5 - 90 kb	28	27
6 - 90 kb	30	29
7 - 90 kb	32	31
8 - 90 kb	34	33
9 - 90 kb	36	35
10 - 90 kb	38	37
11 - 90 kb	40	39
12a - 90 kb	42	41
12b - 90 kb	44	43
13 - 90 kb	46	45
14 - 90 kb	48	47
15 - 90 kb	50	49
16 - 90 kb	52	51
17 - 90 kb	54	53
18 - 90 kb	56	55
19 - 90 kb	58	57
20 - 90 kb	60	59

* ORF-1 of the 90 kb fragment is a partial sequence of the ORF because the 3' end of the ORF, including the stop codon, terminates in the SP6 fragment. The nucleic acid sequence of the 3' end of the ORF-1 sequence, including the stop codon, corresponds to nucleotides 13020-12876 of SEQ ID NO: 103. Thus, the full open reading frame of ORF-1 of the 90 kb fragment consists of SEQ ID NO: 19 (the complementary strand of nucleotides 1635-1 of SEQ ID NO: 1) followed by the complementary strand of nucleotides 13020-12876 of SEQ ID NO: 103.

21 - 90 kb	62	61
22 - 90 kb	64	63
23 - 90 kb	66	65
24 - 90 kb	68	67
25 - 90 kb	70	69
26a - 90 kb	72	71
26b - 90 kb	74	73
27 - 90 kb	76	75
28 - 90 kb	78	77
29 - 90 kb *dptE*	16	15
30 - 90 kb *dptF*	18	17
31 - 90 kb *dptA*	10	9
32 - 90 kb *dptB*	12	11
33 - 90 kb *dptC*	14	13
34 - 90 kb *dptD*	3	7
35 - 90 kb	80	79
36 - 90 kb *dptH*	6	8
37 - 90 kb	82	81
38 - 90 kb	84	83
1 - SP6	86	85
2 - SP6	88	87
3 - SP6	90	89
4 - SP6	92	91
5 - SP6	94	93

6 - SP6	96	95
7 - SP6	98	97
8 - SP6	100	99
9 - SP6	102	101

Table 6: BlastX Results for ORFs in 90 kb Fragment

ORF	Start	Stop	Str	BLASTX (accession numbers, entry title, P-value, E-value)	Polypeptide
1	1637	1	-	emb|CAB88932.1| (AL353863) putative ABC transporter [Strept... 732 0.0 pir||S57562 strW protein - Streptomyces glaucescens >gi|212... 330 e-114 emb|CAB88932.1| (AL353863) putative ABC transporter [Streptomyces coelicolor A3(2)] Length = 593 Score = 732 bits (1870), Expect(2) = 0.0 Identities = 367/462 (79%), Positives = 405/462 (87%)	Type III ABC transporter similar to Streptomyces glaucescens strW gene (resistance to streptomycin); has Walker A, B motifs. Translationally coupled to Orf2.
2	3502	1634	-	emb|CAB88931.1| (AL353863) putative ABC transporter transme. 854 0.0 pir||S57561 strV protein - Streptomyces glaucescens >gi|212 320 4e-86 emb|CAB88931.1| (AL353863) putative ABC transporter transmembrane subunit [Streptomyces coelicolor A3(2)] Length = 623 Score = 854 bits (2183), Expect = 0.0 Identities = 456/637 (71%), Positives = 510/637 (79%), Gaps = 17/637 (2%)	ABC transporter similar to Streptomyces glaucescens strV gene (resistance to streptomycin); has Walker B motif. Translationally coupled to Orf1.
3	4927	3659	-	gi|3913215 1-CARBOXY-3-CHLORO-3,4-DIHYDROXYCYCLO HE 158 1.6e-10 gi|3914351 PUTATIVE 4,5,-DIHYDROXYPHTHALATE DEHYDRO 120 4.6e-06 gi|3913215|sp|Q44258|CBAC_ALCSB 1-CARBOXY-3-CHLORO-3,4-DIHYDROXYCYCLO HEXA-1,5-DIENE DEHYDROGENASE Length = 397 Score = 158 (66.0 bits), Expect = 1.6e-10, P = 1.6e-10 Identities = 59/218 (27%), Positives = 180/218 (82%), Gaps = 24/218 (11%)	Oxidoreductase

4	8364	5410	-	gi| 2506961 D-LACTATE DEHYDROGENASE [CYTOCHROME], MI... 251 5.1e-21 gi|3023651 D-LACTATE DEHYDROGENASE [CYTOCHROME] PRE... 212 1.9e-16 gi|2506961|sp|P32891|DLD1_YEAST D-LACTATE DEHYDROGENASE [CYTOCHROME], MITOCHONDRIAL PRECURSOR (D-LACTATE FERRICYTOCHROME C OXIDOREDUCTASE) (D-LCR) Length = 587 Score = 251 (102.2 bits), Expect = 5.1e-21, P = 5.1e-21 Identities = 119/502 (23%), Positives = 374/502 (74%), Gaps = 91/502 (18%)	Transmembrane, FAD-dependent dehydrogenase
5	8916	8416	-	gi|10803169|emb|CAC13097.1| (AL445503) putative marR-family.. 107 3e-23 gi|15896528|ref|NP_349877.1| Transcriptional regulator, Mar.. 56 1e- 07 gi|10803169|emb|CAC13097.1| (AL445503) putative marR-family regulator [Streptomyces coelicolor] Length = 153 Score = 107 bits (268), Expect = 3e-23 Identities = 66/110 (60%), Positives = 79/110 (71%)	Mar family-related protein Transcriptional regulator Involved in antibiotic susceptibility and resistance
6	9030	10853	+	Gb|AAF67494.1|AF170880_1 (AF170880) NovA [Streptomyces sphe 1017 0.0 emb|CAC13096.1| (AL445503) putative ABC transporter ATP-bin 946 0.0 gb|AAF67494.1|AF170880_1 (AF170880) NovA [Streptomyces spheroides] Length = 635 Score = 1017 bits (2602), Expect = 0.0 Identities = 526/609 (86%), Positives = 559/609 (91%), Gaps = 3/609 (0%)	NovA-related protein (novobiocin biosynthetic gene cluster) that is ABC transporter; has Walker A, B motifs
7	10933	11544	+	emb|CAB91142.1| (AL355913) putative translation initiation ... 64 3e-09 pir||JQ0405 hypothetical 119.5K protein (uvrA region) - Mic... 62 7e-09 emb|CAB91142.1| (AL355913) putative translation initiation factor IF-2(fragment)[Streptomyces coelicolor A3(2)] Length = 835 Score = 63.6 bits (152), Expect = 3e-09 Identities = 74/237 (31%), Positives = 84/237 (35%), Gaps = 6/237 (2%)	Hypothetical protein with no significant match identified by BlastX

8	11990	12850	+	gi|7688708|gb|AAF67495.1|AF170880_2 (AF170880) NovB [Strept 319 2e-86 gi|10803167|emb|CAC13095.1| (AL445503) conserved hypothetic 297 9e-80 gi|7688708|gb|AAF67495.1|AF170880_2 (AF170880) NovB [Streptomyces spheroides] Length = 284 Score = 319 bits (817), Expect = 2e-86 Identities = 156/247 (63%), Positives = 188/247 (75%)	NovB-related protein (novobiocin biosynthetic gene cluster)
9	14038	12878	-	gb|AAF67496.1|AF170880_3 (AF170880) NovC [Streptomyces sphe 520 e-146 emb|CAB71851.1| (AL138667) putative monooxygenase. [Strepto 261 1e-68 gb|AAF67496.1|AF170880_3 (AF170880) NovC [Streptomyces spheroides] Length = 352 Score = 520 bits (1324), Expect = e-146 Identities = 260/346 (75%), Positives = 283/346 (81%), Gaps = 1/346 (0%)	Nov-C related protein that is oxidoreductase
10	14348	14070	-	pir||I39929 hypothetical protein orfM - Bacillus subtilis 78 2e-14 pir||D69817 sulfate starvation-induced protein 6 homolog yg 78 2e-14 pir||I39929 hypothetical protein orfM - Bacillus subtilis (fragment) gb|AAA64350.1| (L16808) Gene disrupted by Tn917 insertion after base 3033.Translation product hydrophilic, no homologues in the databases.; putative [Bacillus subtilis] Length = 372 Score = 78.0 bits (189), Expect = 2e-14 Identities = 37/53 (69%), Positives = 41/53 (76%)	Monooxygenase
11	15697	14522	-	gi|1723069 HYPOTHETICAL 69.5 KDA PROTEIN RV1364C 86 0.04 gi|8928323 SIGMAB REGULATION PROTEIN PHOSPHATASE 2C 85 0.053 gi|1723069|sp|Q11034|YD64_MYCTU HYPOTHETICAL 69.5 KDA PROTEIN RV1364C Length = 653 Score = 86 (37.9 bits), Expect = 0.041, P = 0.04 Identities = 45/153 (29%), Positives = 132/153 (86%), Gaps = 6/153 (3%)	Hypothetical protein

12a	17597	16938	-	gi|728850 GLUCOAMYLASE S1/S2 PRECURSOR (GLUCAN 1,4 113 1.9e-05 gi|138350 GLYCOPROTEIN X PRECURSOR 91 0.0072 gi|728850|sp|P08640|AMYH_YEAST GLUCOAMYLASE S1/S2 PRECURSOR (GLUCAN 1,4-ALPHA-GLUCOSIDASE) (1,4-ALPHA-D-GLUCAN GLUCOHYDROLASE) Length = 1367 Score = 113 (48.4 bits), Expect = 1.9e-05, P = 1.9e-05 Identities = 47/186 (25%), Positives = 158/186 (84%), Gaps = 12/186 (6%)	Hypothetical protein
12b	17870	18682	+	gi|8546911|emb|CAB94663.1| (AL359216) hypothetical protein ... 34 1.3 gi|8546913|emb|CAB94625.1| (AL359215) putative membrane pro... 33 2.9 gi|8546911|emb|CAB94663.1| (AL359216) hypothetical protein SC1D2.05 (fragment). [Streptomyces coelicolor A3(2)] Length = 192 Score = 34.3 bits (77), Expect = 1.3 Identities = 28/94 (29%), Positives = 40/94 (41%), Gaps = 5/94 (5%)	Hypothetical Protein
13	19898	18915	-	emb|CAB94641.1| (AL359215) putative iron transport lipoprot... 250 2e-65 pir||C83282 hypothetical protein PA2913 [imported] - Pseudo... 168 1e-40 emb|CAB94641.1| (AL359215) putative iron transport lipoprotein. [Streptomyces coelicolor A3(2)] Length = 345 Score = 250 bits (632), Expect = 2e-65 Identities = 133/322 (41%), Positives = 188/322 (58%), Gaps = 13/322 (4%)	Iron (ABC) transporter Association with orfs 14 and 15
14	20674	19907	-	emb|CAB94640.1| (AL359215) putative iron transport protein,... 279 3e-74 emb|CAC14366.1| (AL445963) Fe uptake system permease [Strep... 250 2e-65 emb|CAB94640.1| (AL359215) putative iron transport protein, ATP-binding component.[Streptomyces coelicolor A3(2)] Length = 258 Score = 279 bits (706), Expect = 3e-74 Identities = 141/251 (56%), Positives = 181/251 (71%)	Iron transporter Association with orfs 13 and 15

15	21782	20676	-	emb|CAB94639.1| (AL359215) putative FecCD-family membrane t 371 e-102 emb|CAC14365.1| (AL445963) Fe uptake system integral membra 277 2e-73 emb|CAB94639.1| (AL359215) putative FecCD-family membrane transport protein.[Streptomyces coelicolor A3(2)] Length = 368 Score = 371 bits (943), Expect = e-102 Identities = 192/365 (52%), Positives = 248/365 (67%)	Iron transporter Association with orfs 13 and 14
16	23130	21877	-	gi|138350 GLYCOPROTEIN X PRECURSOR 94 0.0088 gi|728850 GLUCOAMYLASE S1/S2 PRECURSOR (GLUCAN 1,4... 83 0.16 gi|138350|sp|P28968|VGLX_HSVEB GLYCOPROTEIN X PRECURSOR Length = 797 Score = 94 (41.0 bits), Expect = 0.0088, P = 0.0088 Identities = 51/216 (23%), Positives = 181/216 (83%), Gaps = 9/216 (4%)	Hypothetical protein
17	23951	23127	-	gi|14591289|ref|NP_143367.1| hypothetical protein [Pyrococc... 46 3e-04 gi|322598|pir||S28604 St12p protein - Arabidopsis thaliana 42 0.006 gi|14591289|ref|NP_143367.1| hypothetical protein [Pyrococcus horikoshii] Length = 248 Score = 46.2 bits (108), Expect = 3e-04 Identities = 31/119 (26%), Positives = 62/119 (52%), Gaps = 2/119 (1%)	Hypothetical protein
18	24966	23953	-	gi|543960 CYSTATHIONINE BETA-SYNTHASE (SERINE SULF 162 4.3e-11 gi|2493892 CYSTEINE SYNTHASE (O-ACETYLSERINE SULFHY... 147 2.4e-09 gi|543960|sp|P32232|CBS_RAT CYSTATHIONINE BETA-SYNTHASE (SERINE SULFHYDRASE) (BETA-THIONASE) (HEMOPROTEIN H-450) Length = 561 Score = 162 (67.5 bits), Expect = 4.3e-11, P = 4.3e-11 Identities = 76/290 (26%), Positives = 243/290 (83%), Gaps = 17/290 (5%)	Hypothetical protein

19	25228	26127	+	gi|8928195 MEVALONATE KINASE (MK) 99 0.00096 gi|8928178 MEVALONATE KINASE (MK) 90 0.011 gi|8928195|sp|Q9V187|KIME_PYRAB MEVALONATE KINASE (MK) Length = 335 Score = 99 (43.0 bits), Expect = 0.00096, P = 0.00096 Identities = 25/61 (40%), Positives = 49/61 (80%)	Hypothetical protein
20	26445	27212	+	gi|731172 SKIN SECRETORY PROTEIN XP2 PRECURSOR (AP... 87 0.019 gi|127749 MYOSIN IC HEAVY CHAIN 86 0.025 gi|731172|sp|P17437|XP2_XENLA SKIN SECRETORY PROTEIN XP2 PRECURSOR (APEG PROTEIN) Length = 439 Score = 87 (38.3 bits), Expect = 0.019, P = 0.019 Identities = 20/54 (37%), Positives = 39/54 (72%)	Hypothetical protein
21	28124	27381	-	emb|CAB56736.1| (AL121600) ABC transport protein, ATP-bindi. .. 351 4e-96 pir||H75293 probable manganese ABC transporter, ATP-binding... 154 1e-36 emb|CAB56736.1| (AL121600) ABC transport protein, ATP-binding subunit [Streptomyces coelicolor A3(2)] Length = 252 Score = 351 bits (892), Expect = 4e-96 Identities = 181/247 (73%), Positives = 193/247 (77%)	ABC Transporter (Mn transporter)
22	28139	29098	+	emb|CAB56735.1| (AL121600) ABC transporter protein, integra... 462 e-129 pir||G75293 probable manganese ABC transporter, permease pr... 208 1e-52 emb|CAB56735.1| (AL121600) ABC transporter protein, integral membrane subunit [Streptomyces coelicolor A3(2)] Length = 283 Score = 462 bits (1177), Expect = e-129 Identities = 241/272 (88%), Positives = 252/272 (92%)	ABC transporter (integral membrane protein) Role in Mn or Fe transport

23	29095	30285	+	gi|6002369|emb|CAB56734.1| (AL121600) hypothetical protein ... 484 e-136 gi|13592175|gb|AAK31375.1|AC084329_1 (AC084329) ppg3 [Leish... 61 2e-08 gi|6002369|emb|CAB56734.1| (AL121600) hypothetical protein SCF76.14c [Streptomyces coelicolor A3(2)] Length = 415 Score = 484 bits (1247), Expect = e-136 Identities = 245/395 (62%), Positives = 287/395 (72%), Gaps = 1/395 (0%)	Hypothetical protein
24	30282	31244	+	gi|6002368|emb|CAB56733.1| (AL121600) putative solute-bindi... 439 e-122 gi|15807666|ref|NP_296243.1| adhesin B [Deinococcus radiodu... 123 2e-27 gi|6002368|emb|CAB56733.1| (AL121600) putative solute-binding lipoprotein [Streptomyces coelicolor A3(2)] Length = 329 Score = 439 bits (1128), Expect = e-122 Identities = 222/315 (70%), Positives = 253/315 (79%)	ABC transporter protein Translationally coupled to orf 23
25	31332	32537	+	emb|CAB56732.1| (AL121600) putative secreted protein [Strep... 620 e-176 gb|AAA59875.1| (M74027) mucin [Homo sapiens] 130 3e-29 emb|CAB56732.1| (AL121600) putative secreted protein [Streptomyces coelicolor A3(2)] Length = 402 Score = 620 bits (1581), Expect = e-176 Identities = 299/402 (74%), Positives = 341/402 (84%), Gaps = 1/402 (0%)	Hypothetical Protein
26a	32816	33427	-	gi|8039818 HYPOTHETICAL 23.1 KDA PROTEIN MLCL581.27 159 5.3e-11 gi|2829591 HYPOTHETICAL 23.0 KDA PROTEIN RV2637 143 4e-09 gi|8039818|sp|Q49642|YQ37_MYCLE HYPOTHETICAL 23.1 KDA PROTEIN MLCL581.27 Length = 214 Score = 159 (66.3 bits), Expect = 5.3e-11, P = 5.3e-11 Identities = 57/197 (28%), Positives = 166/197 (84%), Gaps = 14/197 (7%)	Hypothetical protein

26b	32590	32868	+	gi|15805506|ref|NP_294202.1| penicillin-binding protein 1 [... 33 0.72 gi|7248459|gb|AAF43497.1|AF134579_1 (AF134579) arabinogalac... 32 0.95 gi|15805506|ref|NP_294202.1| penicillin-binding protein 1 [Deinococcus radiodurans] gi|7473266|pir||B75514 penicillin-binding protein 1 - Deinococcus radiodurans (strain R1) gi|6458167|gb|AAF10059.1|AE001907_5 (AE001907) penicillin-binding protein 1 [Deinococcus radiodurans] Length = 873 Score = 32.7 bits (73), Expect = 0.72 Identities = 24/55 (43%), Positives = 28/55 (50%)	Hypothetical Protein
27	34195	35154	+	pir||T36741 probable ABC-type transport system ATP-binding ... 291 6e-78 gb|AAD44229.1|AF143772_35 (AF143772) DrrA [Mycobacterium av... 290 2e-77 pir||T36741 probable ABC-type transport system ATP-binding protein - Streptomyces coelicolor emb|CAB50934.1| (AL096849) putative ABC-transporter ATP-binding protein [Streptomyces coelicolor A3(2)] Length = 332 Score = 291 bits (738), Expect = 6e-78 Identities = 168/303 (55%), Positives = 204/303 (66%), Gaps = 2/303 (0%)	Type I ABC transporter similar to daunorubicin resistance gene, DrrA, in Streptomyces antibioticus; has Walker A, B motifs.
28	35148	36017	+	pir||S32909 hypothetical protein 5 - Streptomyces antibioti... 120 2e-26 pir||T50567 probable ABC-type transport protein, transmembr... 115 6e-25 pir||S32909 hypothetical protein 5 - Streptomyces antibioticus gb|AAA26794.1| (L06249) membrane protein [Streptomyces antibioticus] Length = 273 Score = 120 bits (299), Expect = 2e-26 Identities = 72/226 (31%), Positives = 113/226 (49%)	ABC transporter (integral membrane protein) similar to daunorubicin resistance gene, DrrB, in Streptomyces antibioticus; has Walker A, B motifs.

35	85272	85499	+	pir||T36310 probable small conserved hypothetical protein S... 111 9e-25 gb|AAG29779.1|AF235050_2 (AF235050) CumB [Streptomyces rish... 101 1e-21 pir||T36310 probable small conserved hypothetical protein SCE8.11c - Streptomyces coelicolor gb|AAD18046.1| (AF124138) Cda-orfX [Streptomyces coelicolor A3(2)] emb|CAB38589.1| (AL035654) putative small conserved hypothetical protein [Streptomyces coelicolor A3(2)] Length = 71 Score = 111 bits (276), Expect = 9e-25 Identities = 46/67 (68%), Positives = 56/67 (82%)	Hypothetical Protein
37	86436	87422	+	pir||T36307 hypothetical protein SCE8.08c - Streptomyces co... 175 7e-43 gb|AAA59875.1| (M74027) mucin [Homo sapiens] 94 3e-18 pir||T36307 hypothetical protein SCE8.08c - Streptomyces coelicolor emb|CAB38586.1| (AL035654) hypothetical protein [Streptomyces coelicolor A3(2)] Length = 338 Score = 175 bits (439), Expect = 7e-43 Identities = 120/330 (36%), Positives = 164/330 (49%), Gaps = 13/330 (3%)	Hypothetical Protein Translationally coupled to orf 38
38	87419	88153	+	pir||E83323 hypothetical protein PA2579 [imported] - Pseudo... 102 3e-21 pir||G75588 probable tryptophan 2,3-dioxygenase - Deinococc... 87 2e-16 pir||G75588 probable tryptophan 2,3-dioxygenase - Deinococcus radiodurans (strain R1) gb|AAF12443.1|AE001863_68 (AE001863) tryptophan 2,3-dioxygenase, putative [Deinococcus radiodurans] Length = 287 Score = 87.4 bits (213), Expect = 2e-16 Identities = 73/259 (28%), Positives = 107/259 (41%), Gaps = 37/259 (14%)	Hypothetical Protein Translationally coupled to orf37

Str refers to whether the gene is encoded on the DNA molecule (relative to SEQ ID NO: 1) from left to right (+) or from right to left on the complementary strand.

The BlastX box contains the two top BlastX scores for each ORF (top two lines) and details regarding the database protein entry and the alignment of the ORF to the database entry.

Table 7: BlastX Results for ORFs in SP6 Fragment

ORF	start	stop	Str	BLASTX (accession numbers, entry title, P-value, E-value)	Polypeptide
1	965	1	-	pir||T34645 hypothetical protein SC10H5.07 SC10H5.07 - Stre... 352 2e-96 pir||T36710 hypothetical protein SCH69.11c - Streptomyces c... 206 2e-52 pir||T34645 hypothetical protein SC10H5.07 SC10H5.07 - Streptomyces coelicolor emb|CAA20279.1| (AL031232) hypothetical protein SC10H5.07 [Streptomyces coelicolor A3(2)] Length = 469 Score = 352 bits (904), Expect = 2e-96 Identities = 179/305 (58%), Positives = 216/305 (70%)	Hypothetical Protein
2	989	1948	-	pir||T35566 probable integral membrane protein - Streptomyc... 206 3e-52 gb|AAA53486.1| (U03114) unknown [Streptomyces albus] 139 3e-32 pir||T35566 probable integral membrane protein - Streptomyces coelicolor emb|CAA20393.1| (AL031317) putative integral membrane protein [Streptomyces coelicolor] Length = 315 Score = 206 bits (523), Expect = 3e-52 Identities = 114/311 (36%), Positives = 180/311 (57%), Gaps = 2/311 (0%)	Hypothetical Protein
3	2099	2392	+		Hypothetical Protein
4	3277	2405	-	emb|CAB88937.1| (AL353863) acyl-coA thioesterase [Streptomy... 535 e-151 emb|CAB87210.1| (AL163641) acyl CoA thioesterase II [Strept... 293 1e-78 emb|CAB88937.1| (AL353863) acyl-coA thioesterase [Streptomyces coelicolor A3(2)] Length = 288 Score = 535 bits (1379), Expect = e-151 Identities = 258/288 (89%), Positives = 273/288 (94%)	Acyl CoA thioesterase; enzyme involved in short chain fatty acid biosynthesis

5	5885	3312	-	emb|CAB88936.1| (AL353863) putative helicase [Streptomyces ... 548 e-155 gb|AAG45420.1|AF309494_1 (AF309494) vegetative cell wall pr... 121 1e-26 emb|CAB88936.1| (AL353863) putative helicase [Streptomyces coelicolor A3(2)] Length = 854 Score = 548 bits (1413), Expect = e-155 Identities = 266/323 (82%), Positives = 291/323 (89%)	DNA helicase
6	5963	6754	+	emb|CAB88935.1| (AL353863) putative integral membrane prote... 491 e-138 gb|AAK31375.1|AC084329_1 (AC084329) ppg3 [Leishmania major] 106 2e-22 emb|CAB88935.1| (AL353863) putative integral membrane protein [Streptomyces coelicolor A3(2)] Length = 264 Score = 491 bits (1265), Expect = e-138 Identities = 235/264 (89%), Positives = 246/264 (93%), Gaps = 1/264 (0%)	Hypothetical Protein
7	6850	8403	+	sp|Q9FCB1|DNLI_STRCO PROBABLE DNA LIGASE (POLYDEOXYRIBONUCL... 461 e-141 ref|NP_337667.1| DNA ligase [Mycobacterium tuberculosis CDC... 294 4e-85 sp|Q9FCB1|DNLI_STRCO PROBABLE DNA LIGASE (POLYDEOXYRIBONUCLEOTIDE SYNTHASE [ATP]) emb|CAC01484.1| (AL391017) putative DNA ligase [Streptomyces coelicolor A3(2)] Length = 512 Score = 461 bits (1186), Expect(2) = e-141 Identities = 252/341 (73%), Positives = 267/341 (77%)	DNA Ligase

8	9860	8433	-	emb|CAB93757.1| (AL357613) putative oxidoreductase. [Strept... 299 8e-81 pir||T34726 probable dehydrogenase - Streptomyces coelicolo... 130 9e-30 emb|CAB93757.1| (AL357613) putative oxidoreductase. [Streptomyces coelicolor A3(2)] Length = 481 Score = 299 bits (766), Expect = 8e-81 Identities = 147/185 (79%), Positives = 165/185 (88%), Gaps = 1/185 (0%)	Oxidoreductase
9	10784	9921	-	emb|CAB57411.1| (AL121746) hypothetical protein SCF73.06c [... . 311 3e-84 gb|AAK61383.1| (AY035849) basic proline-rich protein [Sus s... 115 6e-25 emb|CAB57411.1| (AL121746) hypothetical protein SCF73.06c [Streptomyces coelicolor A3(2)] Length = 333 Score = 311 bits (798), Expect = 3e-84 Identities = 166/264 (62%), Positives = 182/264 (68%)	Hypothetical Protein

Str refers to whether the gene is encoded on the DNA molecule (relative to SEQ ID NO: 1) from left to right (+) or from right to left on the complementary strand.

The BlastX box contains the two top BlastX scores for each ORF (top two lines) and details regarding the database protein entry and the alignment of the ORF to the database entry.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

SEQUENCE LISTING

SEQ ID NO: 1

```
   1 GCCACCACCG TACGGCCCTC CAGCACCCGG GCCAGGGAAC GCTCCAGATG ACGGGCGGCC
  61 CGCGGGTCCA GCAGCGACGT CGCCTCGTCC AGCACCAGCG TGTGCGGATC GGCCAGCACC
 121 AGCCGGGCCA GCGCGATCTG CTGCGCCTGG GCCGGGGTCA GCGTGAACCC GCCCGAACCG
 181 ACCTCGGTGT CCAGCCCCTT CTCCAGCGCC TTCGCCCAGC CGTCCGCGTC GACCGCGGCC
 241 AGCGACGCCC ACAGCTCGGC GTCCTTCGCC CCTTCCCTGG CCAGGCGCAG ATTGTCCCGG
 301 AGCGAACCGA CGAAGACATG GTGCTCCTGG TTGACCAGGG CCACATGCTC ACGGACCCGC
 361 TCCGCCGTCA TCCGCGACAA CTCCGCCCCG CCGAGCGTCA CCTCACCGGT GCGCGGTGCG
 421 TAGATCCCCG CCAGCAGCCG GCCCAGCGTC GACTTGCCCG CGCCGGACGG GCCGACCAGG
 481 GCGAGCCGGG TGCCCGGAGC CACGTCGAGC GACACCTTGT GCAGGACGTC GACACCTTCC
 541 CGGTACCCGA AGCGGACCTC GTCCGCCCGT ACGTCCCGGC CTTCCGGGCC GACCTCGGCG
 601 TCGCCCGCGT CCGGCTCGAT GTCCCGGACG CCGACCAGCC GGGCCAGCGA CACCTGGGCC
 661 ACCTGGAGCT CGTCGTACCA GCGCAGGATC AGACCGATCG GGTCGACCAT CATCTGGGCC
 721 AGCAACGCCC CCGTCGTCAG CTGCCCGACC GTCAGCCACC CCTCCAGCAC GAACCAGCCG
 781 CCGAGCAGCA GGACCGCGCC GAGGATCGTC ACGTACGTGG CGTTGATGAC GGGGAAGAGC
 841 ACCGAGCGGA GGAAGAGTGT GTACCGTTCC CACGCTGTCC ATTGAGAAAT CCGCCGGTCC
 901 GACAGCGCCA CCCGGCGGCC GCCGAGGCGG TGCGCCTCCA CGGTCCGCCC CGCGTCCACG
 961 GTCTCCGCGA GCATCGCGGC GACGGCGGCG TAACCGGCGG CCTCCGAGCG GTACGCGGAG
1021 GGGGCCCGGC GGAAGTACCA GCGGCAGCCC ACGATCAGCA CCGGCAGCGC GATCAGCACG
1081 GCCAGCGCCA GCGGGGGAGC GGTCACCGTC AGCGCGCCGA GCAGCAGCCC GGCCCACACG
1141 ACGCCGATCG CCAGCTGCGG CACGGCCTCG CGCATCGCGT TCGCCAGCCG GTCGATGTCC
1201 GTGGTGATCC GGGACAGCAG ATCGCCCGTC CCGGCCCGCT CCAGCACACC GGGCGGCAGC
1261 CCGACGGACC GGACGAGGAA GTCCTCGCGC AGATCCGCGA GCATCTCCTC GCCCAGCATC
1321 GCGCCGCGCA GCCGCATGGA GCGGGTGAAC AGGACCTGGA CGACCAGCGC CACCGCGAAG
1381 ATCGCGGCCG TACGCTCCAG ATGCAGGTCG GTGACCCCGG CCGAGAGGTC CTCGACCAGA
1441 CCGCCCAGCA GATACGGTCC GGTGATCGAG GCGACCACCG CCACCGCGTT GACCGCGATC
1501 AGGACGGTGA ACGCCCTGCG GTGCCGACGC AGCAGACTCC GTACGTAACT CCGCACGGTC
1561 GTCGGTGTGC CCACGGGCAG TGTCGTCGCC GACTCCGGGG CCGCGGGGTC GTACGCCGGG
1621 GGTGCGACGC CGATCATGCC CTCTCCTCGA TTTCCTCGAT GCTCTTCATG GCGGGGACGT
1681 CGCCGCTCTT CATGACGGAG ACGTCGTCAC CGACGCCGTT CACCGCGTCC GCCGCGGCCG
1741 CCTCGTCGTC GGTCTCGCGG GTGACGACCG CCCGGTAGCG CGGTTCGTTG CGCAGCAGGT
1801 CGTGGTGGGT TCCCACGGCG ACGACCGTGC CCTCGTGGAC GAGGACCACC CGGTCGGCGG
1861 CGTCGAGCAG CAGCGGCGAC GAGGCGAACG CCACCGTCGT ACGACCCTGG CGCAGCTTCG
1921 CGATGCCGGC GGCGACCCGT GCCTCGGTGT GCGAGTCGAC CGCGGAGGTC GGCTCGTCCA
1981 GCACCAGCGC CTCCGGGTCG GTGACCAGGG ACCGGGCCAG CGCCAGACGC TGGCGCTGGC
2041 CGCCGGACAG GGACCGGCCG CGCTCGGTGA TCCGGGTCCG CATCGGGTCC CCGTCGTTGT
2101 CGACGGACGC CTGGGCCAGA GCGCTCAGCA CATCGGCGCA CTGGGCCGCC TCCAGCGCCG
2161 TGTCCGGGGT GACCAGGCCC GAGGACGGGA CGTCCAGCAG CTCCTGGAGC GTGCCGGACA
2221 GCAGCACCGG GTCCTTGTCC TGGACCAGGA CCGCCGCTCG TGCGGCGTCC AGCGGGATCT
2281 CGTCCAGGGC GACCCCGCCG AGCAGCACCG ACGGGGTCGC CGCGGCGGCC TTGTCGTCCT
2341 CCTCGCCGGT CTCCGCGTGC CCGCCGAGCC GTTCGGCCAG CCGGCCCGCC TCGTCCGGGT
2401 CACCGCAGAC GACGGCCGTG AACTGCCCGC GCGGAGCCAT CAGCCCGGTC GCCGGGTCGT
2461 ACAGATCACC GGTGGGCGTC ACACCCTCCA CCGTGGCCTC CTGCGCACTG CGGTGCAGCG
2521 ACAGCACCCG CACCGCACGC TGCGCGGACG GCCGGGAGAA GGAGTACGCC ATCGCGATCT
2581 CCTCGAAGTG ACGCAGGGGG AACAGCATCA GGGTGGCCGC GCTGTAGACC GTGACGAGCT
2641 GGCCGACGTC GATGCGGCCG TCCCGGGCGA GCGTCGCCCC GTACCAGACC AGGCAGATCA
2701 GCAGGATCCC CGGCAGCAGC ACCTGCACCG CCGAGATCAG CGCCCACATC CTGGCGCTGC
2761 GCACGGCCGC GCGGCGGACC TCCTGGGAGG CGCGGCGGTA GCGGCCGAGG AACAGCTCCT
2821 CGCCGCCGAT ACCGCGCAGC ACCCGCAGAC CGGCCACGGT GTCCGAGGCC AGCTCGGTGG
2881 CCTTGCCCGC CTTCTCGCGC TGCTCGTCGG CGCGGCGGGT GGCGCGCGGC AGCAACGGCA
2941 GCACGGCCAG GGCCAGCACC GGCATGGCGA GCGCCACCAC CAGCCCGAGG GACGGCAGAT
3001 AGACCGCCAG GCCGACGCAG ATCACCACGA GGGCGGTGGC CGCGGCCGCG AACCGGGAGA
3061 GCGCCTCGAC GAACCAGCCG ATCTTCTCCA CGTCACCGGT CGACACGGCC ACGACCTCAC
3121 CGGCCGCGAC CCGTCGGGTC AGCGCGGAGC CCAGCTCGGC GGTCTTGCGG GCGAGTAGTT
3181 GCTGGACCCG CGCGGCGGCG GTGATCCAGT TGGTCACGGC GGTCCGGTGG AGCATGGTGT
3241 CGCCGACGGC GATCAGTACG CCGAGGGCCA CGATGAGGCC GCCCGCCAGG GCGAGCCGCC
3301 CTCCGGAGCG GTCGATGACG GCCTGGACGG CGAGCCCCAC GGTGACCGGC AGACCGGCGA
3361 TGCCGAGCTG GTGCAGCAGC CCCCAGGAGA GGGACTTCAG CTGCCCGCCG AGCTGATTGC
```

```
3421 GCCCGAGCCA GAACAGGAAG CGAGGGCCCG AACGTACATC GGGGTCGCCG GGATCCGAAT
3481 ACGGAAGGTC GCGAATCTGC ATGACGTCCC AGGGCTCGTG AAACGGAGGT CCGGACAGAC
3541 CTCGAAGACG GGGTGACGTG CAAGGCTCCC TGTTCGTCCC GTTCCGGGGC AACCGGTTTT
3601 TTTCGGTCGC CCCCGCCCTG CGGGGTCCCG GGCCGAGCAG GCCCGGGACC CCACAGACGT
3661 CACTCCGCGG GCTTCTCCGA GTCCATGCCG GACCGGGTCT TCTTCCACTC GCCCCGGGTG
3721 AAGTCCGGGA TCGGCAGGGG CACGCCCTTG GCCTTGATGG ACAGATGGCT CAGCGGCACG
3781 GGGGCCGTCC AGACCGCCGC GTCGTACACG TCGAAGTCGG GCACCAGACC GAGCCGCATG
3841 CACTGCATCA GGCGGAACAC CATGATGTAG TCCATCCCGC CGTGGCCGCC CGGCGGATTG
3901 GCGTGCTCCT TCCACAGCCA GTGGTCCCAC TCGGCGTACT TCTTGAAGTC GTCCCACTGG
3961 TGGTTGGTGT TCGTGGGCTC CAGATAGATC CGCTCCGGGT AGTCCTCGAA CACGCCCTTG
4021 GTCCCGCCGA GGCTGTTGAT CCGCGAGTAC GGGTGGGGCG ACGACACGTC GTGCTCCAGG
4081 CGGATCACCC GGCCCTTGGC GGTCTGCACG AGGCTGATCG TCCGGTCGGC CCCGATGTAC
4141 GACTCCTTCC AGCTCGGGTC GCCCGCAGGC ATGTGCTCCT CGCGGTAGGC GGCGAGGCCC
4201 AGGGGGGTGG TGCCGACACT GCTGATGCTG ACGACCCGGT CGCCCCGGTT GACGTCCATG
4261 TAGTTGGCGA CCGGACCGAA CCCGTGGTTG GGGTAGAGGT CACCGCGCAG CCGGGTGTGC
4321 CACAGCCGCC GCCACGGACC CTCGTAGTAG TCGGGGTCGA ACATCAGCTC ACGCAGATCG
4381 TGGTTGTAGG CCCCGGCGCC GTGCTGCAGC TCACCGAAGA GACCCGCGTG CGCCATCCGC
4441 AGCACCCGCA TCTCGTTCTT GCCGTAACAA CAGTTCTCCA GCTGCATGCA GTGCCGCCGG
4501 GTGCGCTCGG AGAGATCCAC GAGCTGCCAC AGCTCTTCCA GGCGCATCGC GATCGGGCAC
4561 TCCACCCCGA CGTGCTTGCC GTTCAGCATC GCCGTCTTCG CCATCGGGAA GTGCAGCTCC
4621 CACGGCGTCA CCACGTAGAC GAAGTCGATG TCCCCGCGCT TGCAGAGGTT CTCGTAGTCG
4681 TGCTCGTCCT TGGCATAGAT CGCCGGGGCG GGCTGACCGG CGGCCGTCAC CTTCTTGGCG
4741 GCCTTCTCCG CCTTGTCCCG GACCGTGTCG CACACCGCCT TGACCTGGAC GCCCGGGAGG
4801 GCGAGGAAGA GGTCGATCAT GCTGTCGCCG CGGTTGCCGA GGCCGATGAT GCCGACCCGG
4861 ACCGTGGAGC GCCGCTCGAA GGGCACGCCC GCCATGGTGC GGCCCTGCCG GGGAGGGGCG
4921 GCGGCCACGG CTTCCGCGGC GGCGACGGCG TCCGGGGCGC TCCGCCCGGC CGCCGAAGCG
4981 GTGCCTGCGC CCAGTGCGCC GAGGCCGAGT CCGGCCCCGG CCACGCCCGC CGTGGTCCAC
5041 AGCACCGAAC GGCGGCTGGG ATCCTGCCGG TTCACCTCGT CGGCCGCGCC GCTGTGCGGG
5101 GGTATGTCCT GCGGTTCCGG TGCGGGCCGG GCGTCGTCGT TCATCGAGCC TCCAGGTGGG
5161 GTTTGGGGGT TCAGACGGTG CGCGAGCGCG CCCGGTCCCG CCGTACGGAT ACGGGCGGGC
5221 GGGACCGGGG CTCGGTAAGG ACCCTGGAGG GTGAGGCTGA TGGTGCGCAA GGGAAGTATT
5281 TGGACTCTTG TCCTCAAACC TTGGACTTTT CTCACGGCAC GCCGAAGCCC CGACTGGTGC
5341 AACCAATCGG GGCCGTAAAA CGCTCATCTG TGCAGGCCGG CGGGGGTGCC CGCGCCCGCA
5401 GTCACCGACT CACGGGAGAG TCGGCCGGGT GGCGTGTTCC AGTTCGATCA GCGCCGAGCG
5461 GTACGGGTGC CCGGTGGCGC GTTCCATGCC GATCTCGCAC ATCCGGTTCG CCGACAGATA
5521 GGCGTCGTAG GGGCGGCGGT CGACCTCGGC GGCCTCCTTG GCCGTCGCCG AGTCGGTCAA
5581 CTCCTTGTGG AGCATGCCCC GGTCGCCCGC GAACGCACAG CACCCCGCGT CGTCCGGGAC
5641 CACGACCTCC TGCGCGCAGG CCTCGGCCAG CGCCCGCAAC TGCCCCACGT CACCCAGATG
5701 TTCCATCGAA CAGGTCGGAT GCAGGACCGC CGAGCCGGCC GTCCGGAACA CCGTCAGATG
5761 CGGCAGCAGC TCCTCGGCCG CCCACACCAG CGAGTCCACG ACGGTCAGTT CGCGGTGGAG
5821 CGCCCGGTTG TCCTCGGTGA GGTAGGGCAC CACCTCCTCG GCGATGCCGA GCGTGCACGA
5881 GGAGGCGTCC ACGACCAGCG GCAGCGTCCC GCCCGCCGTC CAGCCCCAGG CGGCCTCCAC
5941 GATCCGGTTC GCCATGATCC TGTTGCCCGC GTCGTATCCC TTGGAATGCC AGATCGTCGC
6001 GCAGCACGTG CCCGTGACGT CCTCGGGGAT CCACACCGGC TTTCCGGCCC GCCCGGACAC
6061 GGCGACCACC GCCTCGGCCA GGGAGAGAGC GGGCCCCGCG TCGCCGTCGT CGGGCCCGGC
6121 GAAGATGCGG TTGACACAGG CCGGGTAGTA GACGGCGCTC GCCCCCACGC GTGCGGTGTC
6181 CGGCAGCCGC CGGGCCGCAG CACCGGGGAT CTGCGGCAGC CACTCCGGTA CGAGATCGGG
6241 GCGCACGGCC TTGCGGGCGA GGCGCGTCAC GGCCTGCAGC GGTGCGTCCC CCACCCGGTT
6301 CCCGACGGTG TCGGCCGCCG CCACGGCCAG CCGCGCCGAA GCCTCCACCG CGCGGAAGTT
6361 CTTCGCGGTG AGGGCCGCGA TCCGCTCCTC GCGCGGGGTG TGCCTGCGGT GCCGGAAGCC
6421 CTTCATCATC GCCCCGGTGT CGATGCCGAC CGGGCAGGCG AGTTTGCAGG TGGAATCCCC
6481 GGCGCAGGTG TCCACGGCGT CATAGCCGTA CGCGTCCAGA AGGCCGGACT CCACCGGTGA
6541 GCCGTCGGTC TGCCGCATCA TCTCCCGGCG CAGCACGATC CGCTGGCGCG GAGTGGTGGT
6601 CAGATCCTCA CTGGGGCAGG TCGGCTCGCA GAAGCCGCAC TCGATGCACG GGTCGGCGAC
6661 CGCCTCCACC TTCGGAATGG TCTTCAGGCC CCGCAGATGG GCCCGCGGGT CCCGGTCCAG
6721 CACGATGCGT GGAGCGAGCA CCCCGGCGGG GTCGATGACC TGCTTGGTCC GCCACATCAG
6781 CTCGGTGGCG CGCGGCCCCC ACTCGCGCTC CAGGAACGGC GCGATATTGC GTCCGGTGGC
6841 GTGCTCCGCC TTGAGCGATC CGTCGAACCG GTCCACCACC AGCGCGCAGA ACTCCTGCAT
6901 GAACGCGTCG TACCGGGCCA CGTCGGCCGG CTTCGCCGCG TCGAACGCGA GCAGGAAGTG
6961 CAGATTGCCG TGTGCCGCGT GCCCCGCCAC GGCGGCGTCG AAGCCGTGGC GCGACTGGAG
7021 CTCCAGCAGC GCCGCGCAGG CGTCCGCCAG CCGGGCGGGC GGCACCGCGA AGTCCTCCGT
7081 GATCAGGGTG GTGCCCGAGG GCCGGGAGCC GCCGACGGCG GTCACGAACG CCTTGCGGGC
7141 CTTCCAGTAC CCGGCGATCG TCCCGGCGTC CCGGGTGAAC GCGTTGGTCA CGGACGCCGC
7201 CGGACGCACG AGGTCCAGAC CGGCCACGAC CGCGTCCGCC GCCCGCTCGA ACGCCGCCCG
7261 GCCCGCCTCG TCGGCCGCCC GGAACTCCAC CAGCAGCGCG GTCGTCTCCC GGGGCAGCGC
```

```
 7321 CGCCCAGTCC GCCGGAACGC CCGGCACGCT GACGGAGGCG CGCAGGGTGT TGCCGTCCAT
 7381 CAGCTCCACG GCGATCGCCC CCGCCTCGTT GAACCGGGGC ACGGCCGCCG CGGCGGCGGT
 7441 GAGGGAGGGG AAGAACAGCA GGCCGCTGGA GACCCGCCGG TCGAGCGGGA GGGTGTCGAA
 7501 GACGACCTCG GAGATGAAGC CGAACGTGCC CTCGGAGCCG ACCATCAGCC CGCGCAGGAT
 7561 CTGCACCGGC GTCGCCCCGT CGAGGAAGGC GTCCAGGCGA TAGCCATTGG TGTTCTTGAT
 7621 CGTGTACTTG GCGCGGATCC GGGCGGTCAG CTCCGCGTCC GCCTCGATCT CCGCCTTCAG
 7681 CTCCAGCAGC CCCGCGCACA GCTCCGGTTC GGCGTGGGCC AGCTCCTCGT CGGCGGCGGG
 7741 GTGCGCGGTG TCGACGACGG TGCCGCTCGG CAGCACGAAG GTGAGCGAGG CGAGCGTGCG
 7801 GTAGGAGTTG CGGGTGGTGC CCGCCGTCAT GCCCGAGGCG TTGTTGGCGA CGACCCCGCC
 7861 GAGGGTGCAG GCGATGGCGC TGGCCGGATC GGGGCCCAGC AGCCTGCCGT ACCGGGCGAG
 7921 GGCGGCGTTG GCCCGCATGA CGGTGGTGCC CGGCAGGATC CGGGCCCGCG CCCCGTCGTC
 7981 CAGCACCTCC ACGCCGGTCC AGTGACGGCG TACGTCGACG AGGATGTCCT CGCCCTGGGC
 8041 CTGGCCGTTG AGGCTGGTGC CCGCGGCCCG GAAGACCACG GATCGGCCCT TGCCATGGGC
 8101 GTACGACAGG ATCGCGGACA CGTCGTCGAG GTCCTCGGGG ACCAGCACGA CCCGGGGGAG
 8161 GAAGCGGTAG GGGCTGGCGT CGGAGGCGTA CCGCACGAGG TCGGAGATCT TCCAGAGCAC
 8221 CTTGTCCGCG CCGAGCAGCG CGGTCAGCTC GCTCCGCAGC GGCTCCGGGG TGCCGCCCGC
 8281 GCTGCCGTCG GTGACCCGGT CGGGGGCGGG TTCCCGCGCC GTTCCGGGGC GCAGCGCTTC
 8341 CGGGTCGGGC TCCAGCAGCG GCATGTCGGC CTTCCCCTCG GCTCGGCGCT CAGCGGTGGC
 8401 ACGCGGCAGC GGCGCTCAGC AGTGGCGCTC CGGCATTCCG TCGACCAGAG CGGACAGCAG
 8461 CTCGCCGAAC ACCTCGCGCT GATCGGCGGT CAATGGAGCC AGGATCTCCT CTGCGGCGGC
 8521 CCGGCGCGCG CTGCGCAGGG ACCGCAGCGT GGCGCGCCCC TCGTCGGTGA TCTCGATACG
 8581 GACCACCCGG CGGCTGTCGG GATCCGGGGC GCGGCGCACC CGGCCGCTCG CCTCCAGGGC
 8641 GTCGACCAGC GTGGTCACGG CGCGCGGGAC GACGTCGAGC CGTCGGGCCA GATCCGCCAT
 8701 CCGGGGGGCC GCGTCGTAAC TCGCGACCGT CCGCAACAGG CGGAACTGGG CCGGAGTGAT
 8761 GTCGATCGGC TCCAGCTGGC GGCGCTGGAT GCGGTGCAGC CGCCGGGTGA GCCGCAGCAG
 8821 CTGTTCGGCG AGCAAGCCGT CACGGGAGTC CCGGGAATCG CGAGAGTCCC GGGACTCGGG
 8881 GGAATCAGGG GAGTCGGGGG AATCCGGGGC GTCCATACGG GAACAATATC AGGACCTTGT
 8941 TCATTGTGAG CATAGGTAAC AATGAGCTAG GCTCTCACTG TGCGGGACCG GGACTGCCCG
 9001 GCCCCGCCTC ACGCCCGACG AAGGAGCCCA TGAAACCCGA CGAACCCACG TGGACGCCCC
 9061 CGCCCGATGC CCGCCCCGCC GCCGACCGGC GGCCCGCCGA GGTGCGCCGC ATCCTCCGCC
 9121 TCTTCCGCCC CTATCGCGGC CGCCTGGCCG TCGTCGGCCT GCTGGTCGGC GCATCCTCCC
 9181 TGGTCGGGGT CGCCTCCCCG TTCCTGCTGC GCGAGATCCT CGACACCGCC ATCCCGCAGG
 9241 GACGCACGGG CCTGCTGACC CTGCTGGCGC TCGGCATGAT CCTCACCGCC GTGATGACCA
 9301 GCGTCTTCGG CGTGCTCCAG ACCCTCATCT CGACCACCGT CGGCCAGCGC GTCATGCACG
 9361 ACCTGCGCAC CGCCGTCTAC ACCCAGCTCC AGCGGATGCC GCTCGCCTTC TTCACCCGGA
 9421 CCCGCACGGG CGAGGTCCAG TCCCGCATCG CCAACGACAT CGGCGGCATG CAGGCGACGG
 9481 TCACCTCCAC CGCGACCTCG CTGGTCTCCA ACCTCACGGC CGTCATCGCG ACCGTCGTCG
 9541 CCATGCTCGC CCTCGACTGG CGGCTCACCG TCGTCTCGCT GCTCCTGCTG CCGGTCTTCG
 9601 TCGCGATCAG CCGCCGCGTC GGCCGGGAAC GCAAGAAGAT CACCACCCAG CGCCAGAAAC
 9661 AGATGGCCGC GATGGCCGCC ACCGTCACCG AGTCCCTCTC GGTCAGCGGC ATCCTCCTCG
 9721 GCCGCACGAT GGGCCGCTCC GACTCCCTCA CCCAGGGCTT CGCCGAGGAG TCCGAGCGCC
 9781 TGGTCGACCT CGAAGTGCGC TCCAACATGG CCGGCCGCTG GCGGATGTCC GTGATCGGCA
 9841 TTGTGATGGC CGCCATGCCC GCCGTCATCT ACTGGGCGGC CGGACTCACC TTCGCGTCCG
 9901 GAGCCGCCGC CGTCTCCATC GGCACACTCG TCGCCTTCGT CACGCTCCAG CAGGGGCTGT
 9961 TCCGCCCGGC GGTCAGCCTG CTCTCCACCG GTGTGCAGAT GCAGACCTCC CTCGCCCTCT
10021 TCCAGCGCAT CTTCGAATAC CTCGACCTCA CGGTGGACAT CACCGAACCG GAACACCCGG
10081 TCCGGCTGGA GAGGATCCGC GGCGAGATCG CCTTCGAGGA CGTCGACTTC AGCTACGACG
10141 AGAAGAACGG CCCGACGCTG ACCGGCATCG ACGTGACCGT CCCCGCGGGC GACAGCCTCG
10201 CGGTCGTCGG CTCCACCGGC TCCGGCAAGT CCACCCTCAG CTACCTGGTG CCCCGGCTGT
10261 ACGACGTCAC CGGCGGCCGG GTCACGCTCG ACGGCATCGA CGTCCGCGAC CTGGACTTCG
10321 ACACCCTCGC CCGGGCCGTC GGCGTCGTCT CCCAGGAGAC GTACCTCTTC CACGCCTCGG
10381 TCGCCGACAA CCTCCGCTTC GCCAAGCCGG AGGCCACCGA CGAGGAGATC GAGGCCGCGG
10441 CCCGCGCCGC GCAGATCCAC GACCACATCG CCTCCCTGCC CGACGGCTAC GACACGATGG
10501 TCGGCGAGCG CGGCTACCGC TTCTCGGGCG GCGAGAAGCA GCGCCTCGCC ATCGCCCGCA
10561 CCATCCTGCG CGACCCTCCG GTGCTGATCC TCGACGAGGC GACCAGCGCG CTCGACACCC
10621 GTACGGAACA GGCCGTGCAG GAGGCGATCG ACGCCCTGTC CGCCGGACGG ACCACGCTCA
10681 CCATCGCGCA CCGGCTCTCC ACCGTCCGCG ACGCGGACCA GATCGTCGTC CTGGAGGACG
10741 GCCGGGTCGC CGAGCGCGGT ACGCACGAGG AACTGCTCGA CCGCGACGGC CGCTACGCCG
10801 CCCTGATCCG CCGCGACTCC CACCCGGTCC CGGTCCCGGT CCCGGCTCCC TGACCACCCT
10861 TGTCGGGCCG GCCCTCGATC AGACCGCCCC TGACGTCACC GCCATGGCCC GCATACGGCA
10921 TGATCGCCGC GCATGAGAGC TCTCCTCGGG GTGGAACTCC CCGGCTACCG CACCGTCGAC
10981 ACCGACACCT GGCTGAACGA CCACGGCGAT GTGCTGTCCT TGCACTTCTT CGACCTGCCG
11041 CCGGACCTGC CGGCCGCGCT GGACGACGGC CCGGCCCTGC GGCACGGTCT GACCCACTTC
11101 ACCGCCAGGG CGGGCGGCGG CCTCATCGAG ACATCGGTGA AGCGGCTGGG CGAGCTGCCC
11161 GCCCTGCGGC AGATACTCAA ACTGCCGCTG CCGAACCAGC CCAGCGGCCA GGCGTTCATC
```

```
11221 GGCAGCTTCA CCGTGCCGCG CGCCGGATGC AGCACCGTGG TGAAGATCCA GGCGGCGGAG
11281 CGCGGCATGA CGGGCATGCG GGAAGCCGTG GTGATGGCCA AGCTCGGCCC CGACCAGTAC
11341 TTCCGGCCGC ACCCCTACGC CCCCGAGGTC CAGGGCGGGC TGCCCTTTCA CACGGCGGAT
11401 CACGTCCAGT GGGACGCGGA GTTCCCGGAC CATCCGCTCA CCCGGGTCCG CCGGACGCTC
11461 GACACCCTCG CGGCGGCGGT GACGGTGGCA CCCGAGTTCG CCGCGCTGCC GCCCTTCACC
11521 GGACCGGCTC AGGCGAACGG CTGAGCCGAC CGGCTGCGTA CACACAGCAC ACAGCACACA
11581 GGGCACACGG CGCACACAGC ACACACGGCG GCGCCGCCGC TCCCGTGGGA CGGGGAGCGA
11641 CGGCGCCGGG CGGAGCAATG GTCAGACGAG CCAACCCACG AAGTGGACGA CGCCGGCAAG
11701 CAGGTTGGTC AGGAAGTTCA TCTGGTCTTT CTCCTTGTAC GTGGTGCATC TGTGGGACTG
11761 CGCAGTAGCG GTCTGCAGCC CGTTGACTGC GCTCTGCAAT CATCACGCCC CGGACGAGTG
11821 AAGAGCAACG AATCCCCTGA CGATCACGCG TTCCAGCGAA CACCCGATCT CTTGTTCGTG
11881 TGTTCCGGCT ACGGGTGTTC TGTCCGCGTC GTACGGCGTT CGTGTCGCCG GGGCCGACGC
11941 CGTGGTCGGG CTACCGGCCC TGGCTCGCAC CCCGGGTTAA CGTGCCCGCA TGGTGAACGA
12001 GTCCCCGGAC GCCCGACCCC GTCGCAGACT CCGCCCGACC CGCCGCGGAA AGATCGTCCT
12061 GGTCGTCGGC GCACTGCTCG TCGTGACGGC CGCCGTCCTG ATCCCCCTGT CCCTGACCGG
12121 ATCGGACGAG CCGCCGAAGA AGCAGGAGAC CCCGCAGAGC ACGCTGATGA TCCCCGAAGG
12181 CCGCCGAGTG TCCCAGGTGT ACGAAGCGGT CGACAAGGCG CTCGACCTGA AGCCCGGCAG
12241 CACGCTGAAG GCCGCGTCGA CGGTGGACCT GAAGCTGCCC GCCCAGGCCG AGGGCAACCC
12301 CGAGGGGTAC CTCTTCCCGG CCACGTATCC GATCGACGAC ACGACCGAGC CCGCGGGCCT
12361 GCTGCGCTAC ATGGCCGACA CCGCCCGCAA ACACTTCGCC GCGGACCATG TCACGGCCGG
12421 GGCCCAGCGG AACAACGTCT CCGTCTACGA CACGGTCACC ATCGCCAGCA TCGTCCAGGC
12481 CGAGGCCGAC ACCCCGGCCG ACATGGGCAA GGTGGCCCGC GTCGTCTACA ACCGGCTGCT
12541 CAAGGACATG CCGCTCCAGA TGGACTCCAC CATCAACTAC GCCCTCAAGC GCTCCACCCT
12601 GGACACGTCG ACCGCCGACA CCCAGCTGGA CAGCCCGTAC AACAGCTACC GGATCAAGGG
12661 CCTGCCGCCG ACGCCCATCG GCAATCCGGG AGAGGACGCG CTGCGCGCCG CCGTCAGGCC
12721 CACGCCCGGC CCCTGGCTCT ACTTCGTCAC GGTCGGCCCC GGCGACACCC GGTTCACGGA
12781 CAGCTACGAC GAGCAGCAGA AGAACGTCGA GGAGTTCAAC CGCGGCCGTG GCTCCGCCAC
12841 GACGGGCTGA CCGAATCGGC AGACGGGGCG GGGGGATTCA CACCCCCGGC ACGGGCGCGG
12901 GCACGGAGAC GACCGCCGAG GCCCCTCCGT CGGCGCCCGT CTCCTTCAGC AGCCGCATGA
12961 CCGACCGGAC CGCCGCGCGG CCGGCGCGGT TCGCGCCGAT GGTGCTGGCG GAAGGGCCGT
13021 ACCCGACGAG ATGGACGCGC CCGTCCCGTA CGGCACGGGT GTCCTCGGCC CGGATGCCAC
13081 CACCCGGCTC GCGCAGCTTC AGCGGGGCCA GATGGTCCAC GGCGGGCCGG AACCCGGTCG
13141 CCCAGAGGAT CACGTCGGTC TCGACGGTAC GGCCGTCGTC CCAGGCCACA CCGGTCGGCG
13201 TGATCCGGTC GAACATCGGC AGCCGGTCCA GCACTCCCCG CTCCCGGGCC CGCCGCACGG
13261 CATCGTTCAG CGGCAGCCCG GTCACGCTGA CCACGCTCTT CGGCGGCAGC CCGTTGCGTA
13321 CCCGCTCCTC CACCATCGCC ACGGCCGCCC GCCCCCACTC CTCGGTGAAC GGACCTTCGC
13381 GGAACACCGG TTCGCTGCGG GTCACCCAGA AGGTGTCGGC CGCGTGCTCG GCGATCTCCA
13441 TCAGATGCTG CGTACCGGAA GCGCCACCCC CGACCACGAG GACGCGCTGC CCGGCGAACT
13501 CCTCGGGCCC CGGATAGTTC GCCGTGTGCA ACTGCCGCCC CCGGAACGTC TCCTGGCCCG
13561 GATAGCGCGG CCAGAACGGC CGGTCCCAGG TGCCGGTGGC GTTGATCAGA GCCCGCGCGG
13621 CGTACGTCCC CTCGGACGTC TCCACCAGCA GCCGACCGCC GCTTCCCTCC CGTACGGCGC
13681 TCACCTCCAC GGGCCGGTGG ACCCGCAGGC CGAAGCGGTC CTCGTACGCG GCGAAGTACG
13741 CGCCGATCAC CTCCGACGAG GGCCGGTCGG GGTCGGCCCC GGTCAGCTCC ATGCCCGGAA
13801 GCGCGTGCAT CCCGTGGACC TTGCCGTACG TCAGCGAGGG CCAGCGGAAC TGCCACGCAC
13861 CGCCCGGCCG GGGCGCGTGG TCCAGCACGA CGAAGTCGTT GTCCGGCTCC AGCCCGACGC
13921 GGCGCAGATG GTAGGCGGCG GACAGTCCCG CCTGACCCGC GCCGATGACG ACCACGTCCA
13981 GCTCGCGCAC CCCAGAAATG TTCACGCTTC TACTAACTCG TCGGGCGCCC GGGATCATCC
14041 CGGGCGCCCG ACGAGCGTCA CCGCACGGCT CAGCGACCCC CGGCGAGCAG CAGGGGAGCC
14101 CCGCCCGGCG CCGTGGCGGT CCGGCTCTCG GCGCCACTCA CACCCAGCAG CGGCGACGCG
14161 GGCACGGTCG ACAGCAGCCC GCGGCGGGAC AGCTCGGGTG TGACGCCCTC GCCGAACCAG
14221 TACGCCTCCT CCAGATGCGG ATATCCGGAG AGCACGAAGT GCTCCACGCC CAGCGCGTGG
14281 TACTCCTCGA TCCGGTCCGC GACCTCCGCA TGGCTGCCCA CCAGCGCGGT CCCGGCCCCG
14341 CCGCGCACCA GACCGACCCC TGTGTGCGCG AGTGGAGATG ATCCTGCACG GCACGCCCCA
14401 CATGCGATCG AGCGTTTCTG AGGTCTGTGC CGTGGCCTCT CGGACATGGC CACGGCGAGG
14461 GAATCGGTGT TGCAGAGTGA CGTGTGTCGT TGAGTCGGCG GCTCGGGCGG GGGTGGCGTG
14521 GTCAACGGTT CGGTGGGCGG GCGGCTGAGG CTTGGTGGGT GTCCGCGCCG GAGCTGACGT
14581 GGCGGTGTAC TTCTTGTGGG CCGTGCCAGT CCAGGCACAC CACGGTGGCG TCGTCTTCGA
14641 GGCGGCCGCC TGCGGCGTCG CGTACGGCGG AGGTCAGCAT CAGGGTGGTC TCGCGTGGGT
14701 GCAGGCTGCG GGTCTGCCGT AGCAGGGCAG CTACGTCGAT CTTCTCTCCG TGGCGTTCGA
14761 GCATGCCGTC CGTCAGCATG AGGAGCCGGT CTCCCGGGTG CAGGTCCAGG GTTTGGACGC
14821 GGTAGGGGCG GGGCGAGACG ACGGCCAGCC CGAAAGGCTG GTCGACCTGG CAGGGGATGG
14881 TTTCCACCAT GCCTGCACGC ATGCGCATGG GCCAGGGATG GCCGGCGTTG ACGAGCTCGG
14941 CCTTTCCGGT GTGGAGGTTG ATGCGCAGCA GTTGTCCGGT GGCGTGGCCC TGTCCGTGGC
15001 TGGTCAGGGC CTGGTCGCCC TGGCGGGCCT GTTCGGCGAG GGGGGCTCCG GCGCGGCGGG
15061 CTCTGCGCAG GGCGCCCACC AGGACGGTGG CCGCCAGGGC TGCGCCGAGG TCATGGCCCA
```

```
15121 TGGGGTCGGT CACCGACAGG TGCAGGGTGT CGCGGTCCAG TGCGTAGTCG AACGTGTCGC
15181 CGCTGAGGTC CTCGGAGGGC TCCAGGCTCC CGCTCAGGGT GAACTGCGCG GCCTCGCAGG
15241 ACAGGGCCTG TGGAAGCAGC TGATACTGGA TCTCCGCTGC CAGGGTCGGG GGTCTGGAGC
15301 GTTTGCCCCA GGTGTAGAAG TCGGTGAAGC GCCCGTTGGC GATCACGACG TAGGCCAGCG
15361 CGTGAGCGGC TTCCCCGACA GCGAGCACAA CCTCTTCCTC GTCGCTCCTG CCGGCCGGCA
15421 GGAGCAGTTC GAGCAGACCG ATCGCGTCCC CCCGGTTGGT CACGGGAACT ATTACCCGCT
15481 GTTCTTGTCC GGCGGGCTCG TGATGCGGCC GCTGGGTGCG GATCACCTGC TCGTAGACGC
15541 TCCCCCCGAA CAGAGGGATC CGCTCCGTTT CGTTTTCACT GCCCGCGGCA GTCGTGGTGG
15601 AGAGCCGCGC GAGCGCTCTA CCGGTCAGAT CCACAATCAG GAATGTGACC TTCGTAGCCG
15661 CGAACCGCCT GCGCAGATCT TCTGCGACCA CGGCCACGGC CTCAACCGGA GCCGCCGTCT
15721 CCGCCGCCGT CAGCAGTCGG GACAGGTCGC TCGAGCCACG GCTCATGGCA GCGGTTCCCT
15781 TCTCTGGATG TTTGGGGCCC GTTGCGCCCC GCCGCCCAGT CGCCCCTCCT CGTACCTGCC
15841 TTGCGCTCCA CGGTGGTCGA CAGCGAGCAG CCCGGCACGG GCCCGCCCGC TCTCGGCCCA
15901 TTGCGTGACA GTCCTGCATC CACCTGTTCC AGTCTGAACC TCAATCGGCC CTTTGTCCGG
15961 ATGAGGGACC GGGTCGGCCG GAGGCGAGGC GCCACCGGGT GAGGAAGGCG CCGACCGCCA
16021 CCTCGATGGG GTCGGGGCGG ACGATGTCAC CGAACTCGGT GGCGCTTCTC CCGCACCACC
16081 CCGAGCAGAG CGTTGTCCAC GGGCGATGTT CGTCGCCGAC ACCAGCAGGA CACGCCGCCC
16141 CTGGGCGATG CGGTCGCCGA TGGCTCGCCG GAGGACGGTG GTCTTCCCTG TACCAGGCGG
16201 CCCCACACCA GGTGGACGCC CTCGCCGAGG CATGCTCGAT ACGCGAGCCT GGGCAGGATG
16261 GAAGCCCGGC GGATCGATGG CGTGGGCAGA GCGACCGCCG ATCATCGCGG TCGCCAGAGC
16321 AGTGGCGAGA GGATGCTCAC CCAGCCAGGC GATACCGTCA CGCAAGGCCT CGATCAGGAA
16381 GGTGGGCGGC TGCTTGAGCA TCCACAGGTG AGGGTCGTCG ATCTGCGAAA TCTGCTACCC
16441 GCAATGTGAG CAGGGAGCCG TTCTGTACAG CCTCGAAGAC TGCGAAACCC TCGATCTCGA
16501 TGCCGTCGTT GCCCGTCCCG GGCGGCCTCA GGGAGTCCAG CTGCCCAGGT CCGATGTCGG
16561 AGCCCAGCAG ATCGACCACG TACCGGCCCG GGTCACCGCT CCTGGCCGCC CGCCCGACGA
16621 GCTGCCAGCG TGGCTGCCTG CCCACGCCTC CTTCGACAGC GATCCACTCT CCGAGTGCCG
16681 AGGCGATTTT CTCACGCCAT CCCACCTACC GTCCCCCCGA TCAGCCTCGG TCCGATCGCC
16741 TGCCCGCTGC TGCGCTGTGC CCTCCGGCTG CGATCCGGTT CGCTCGAAGT GCCTGCGGCC
16801 TGTTCACGGG GCCGGTGGAT CCGCTCCGGA TGCGCTGTCC TTGCAGGCAC GTTCGTCCAG
16861 GCAGTCGGCT CCCGAAGCCG TCCAGGGCGC ATCACTCCGC AGGGAGCTAG AGGGCTGTCC
16921 CGTAAAAAAC CTCCGTCTCA GGGGCGTTGG GGGTAGTCAG GGTGATCTGC GTAGGGTGAC
16981 GCGAGACCGA GCAGGTCATC GCATGGCCAG GTCGCGCCGT CGTACCGACA GTCATCGCGG
17041 TTGTCCCACG GCAGTTCGGG GTCACCCGTG GCAGGGCTGA GCCCATGTCG GGCGAGCACG
17101 CGCCGCTTGG CCTCGGCTTC CCGCAGGACG CGAGCCGGGT CGTGCAGCGC GACGTGCAGG
17161 GCGATCGTGG AATGGAAGCC GGAGAGTTCT CCCTGGCAGA AGTCGACCGT GTGTCCGTGA
17221 GGGGCCCACT CCCCGCAGCC GTCACCGTCG CACCGGCCGG CCAGGTTGGC CTCCTCGTCC
17281 AGCCGAGCGT GGAGGAACGT CACAAGGTCT TGGCTCATGG GGTCATCCTG GCCGACGGCT
17341 CGGCCGGTGG CCGGCCCACT GTTTGCGAAC TTGCGGGCGG TCTGTCGCAG GGCACCGCCC
17401 TGTCCGTGTT CGGCACGGAC GCGGGAGCGG GAGGCCCCTT GGAACGCGAA TGCTCCAGCT
17461 TCGAAGGCAA CGGCGAGCAG CAAGGGGTCC GGCACCATCC CCGCACTCCG TGCGCCACAC
17521 CTGCGCCTCT CGCGCTCTTG TGTCACGAGA ACTCCCAGAC CGCAAAGCGC CACACCCACC
17581 TGCAGTCGGA CGCGCATAGT CCGCCCTGGA TGCTGCGGAA TCGCATCCTC AGCCCGCGTC
17641 GCAGTATAGA TAGTCCGGGT TGTCCCAACG GTTGTTGGTG GACACCAACA AGGACAATGC
17701 CAACCGTCAG GTTCGGCTGA GCTTCTTCGG AGGGTGAGCC GATCTGGTTG CACGAGAGGC
17761 GGGAAGTCCG GCCGCACCAA CCGGACGCTC GGACCGTGTG TCCGAACCTG TCACGCAAGC
17821 CTGCACAGGC CCCCGCCGTG CAATCGAAGG GCTGCTCCTG GTGCCCGTGA TGTCGGTACG
17881 CGATCTCGTC GGGATGCCGT GTCACCCGTG CGAACCGCCA CGCCGCGCCG AGGGGCGCCG
17941 GCGCGGCGTG GGGAGGATGA GGTGGTGGAA GGGGGTGCTG ATGACGGTTC GGCATCAGGG
18001 GGTGCGGTGG TGGTTCGCTC TTCTCGCTCT CGTCGGGTGC GTGGTCTGTG TCCTCTGCGT
18061 CGTCGCGCTC AGCGGGGCGG GGCACTACTT CGGGCTCTCC TTGTGGGCGG GCATCGCGCT
18121 CGTGGTGGTG GGGGCGCTGT TTCCCCTCGG GGGGCTGGGC TTCCTGTACT GGGTGGACGA
18181 CGGCCGGTCC GAGGACAGCT TCCTCGTGAA GTTCCTGTGC TTCGTCGCCC ACTCCGCCGT
18241 CCTCGGGCTG GCAGCCGTCT CGTGCACCGG GGCTGAGGCG TGGGCCTTTG AGCAGCGCGG
18301 GCGGTGGACG GAGGCGACGG TCGTGGGATA CAGCCCGCCC CGGGTGGTCC CGGGTGATCC
18361 GCCGACGAAG GTGCGGGCGT CCTGCGCGCT GGAGACCGCC GAGGGCGAAC GCGTCCGGCC
18421 CCGGCTGCCG GAGGGCCGCG GCTGCCGCGA CGGGGTGCGG CACGGGTCCC GCCTCGACGT
18481 GCTGTACGAC CCCCGGGGTC TGCTGGCGCC CCGGGCCACC GAGCCCATGG ACCACGGCGT
18541 CACCGTCCCG GTCCTCGGGG GCGTGGCGAC CCTGTCCGGT TTCCTCGGCT GTGTCGCCCT
18601 CGCCTGGCGG TGGGAAACCC TCCGGGTACG CAGCGCGCGC CGCACGGCAG CGCGCCGAGG
18661 GCGGGAATCC GCAGCCGGTT AGGGGGTGGG GGCGTTCGCC GGCTCTCCTT GCCGCCGTGA
18721 CCTGGAGCGC GGCGCGGCTG GAGCCCACCT GCCGGGCCGA GTAGTTGCCT GCACTGCGCC
18781 CTTCTCGCCG TGGGAGATCG TGGCTGAGGC GATGGGCGGA AGACACCCGG CCTTCCCCGG
18841 TTCAGAGGGG AAGGCCGGGT GTCAGGCGCA AGGACCTGCG AGAACCCGGA AGGATCCTGC
18901 TGCCGGGCCG GTCATCATTT CTTGAATGCG CGCATGTACT TTCCGAACTT CTCCAGGCCG
18961 TCGATATTGC GCGGGCTGCT GATGCCCTCG TTGTAGTCGA GGACGAAGAA GTTGTTCTTC
```

```
19021 TTCACGGCCG GCAGTTCCTT GGTGTGCGGC GACTTCTTCA GGAACTCGAT CTTCTTCTCG
19081 GCGGGCTGGT CGCCGTAGTC GAAGATCATG ATGACCTCGG GCTCGGCCTG GGTGACGGCT
19141 TCCCAGTTCA CCTGGGTCCA GCGCTCCTCC AGGCCGTCGA AGATGTTCTT CCCGCCCGCG
19201 GTCTTGATGA TGTCGTTGGG CGGCACCTGG TTGCCCGCCG TGAACGGCTG GTCGGTCCCG
19261 GAGTCGTAGA GGAACACGGG CACGGGCTTG CCCTTCGGAG CCTGCTCGGC GACGGCGGCC
19321 TCGCGCTTCT TCAAGCCGGC GACGACCTTC TCCGCCTCCT CTTCGACCTG GAAGATCCGT
19381 CCGAGGCGTT CGAGGTCGGT GTAGAGGCCC TTGAAAGGCG TCAACTTCTC CGGATGGCCC
19441 GGGTAGTTGT AGCAGCTCTC ACTGTGCATG AAGCTCTGTA CGCCGAGCTT GTCGAGGATC
19501 TCCGGGGTGA TGCCCCGCTG GTCGCTGAAG CCCGAGTTCC AGCCGGCGAC GACGAAGTCC
19561 GACTTGGCGT CCACGACGAT CTCCTTGTTG AGGAGGTCGT CGCTGAGCAT CTTCACCTTG
19621 GCGTAGTCCT TCGCCCAGGG AGACTCGCTG ACCGGCGGGT TGGCCGGCGG CATGACGTAG
19681 CCGTGCACGT GGTCGGCCAG GCCCAGACTG AACAGCTTGT CGGCGCTGCC GCCCTCGTAG
19741 GCGACGGCCC GCTTCGGCAC CGTGTACTCG ACGGACTCGC CGCAGCGCTT CACGGTGCTC
19801 TTCCCGGAGC CCTTGCCCTG GGATTCGACC TCGGCGCCAC ACCCCGTGAG CAGGAGCGCG
19861 GACGCGGCGA CGGGGATGGC GAGTTTGGTG AACTTCATGG TCTTCCTCAG GAATCGAGTG
19921 AGTAGAGCAA CTGGGGGTCG CCCGTCAGCG GATGCGGGAC GACGGAGGCG CGGACCCCGA
19981 ATACCTCGTC GACGAGTTCG GGCGTGAGGA CGTCCTTGGG CGTGCCCGAG GTGATCAGGC
20041 GGCCTTCGCT GAGTACGCCG ATCCGGTCGC ACGCGGCGGC CGCGAGGTTC AGGTCGTGGA
20101 GTACGACGAG GACGGTCAGG CCGGCACCGC GCAGCAGGGA CAGGAGCCGC ACCTGATGGC
20161 GTACGTCGAG ATGGTTCGTC GGCTCGTCGA GGACGAGGAT CTTCGGCTCC TGCACGAGGG
20221 CGCGGGCGAG CAGGACGCGC TGGCGCTCGC CGCCGGAGAG GGTGAGGATG CCGCGTCGGG
20281 CCAGGTGCAG GATGTCGAGC CGACGCATGG CGTGCTCGCA CAGATCCCGT TCGTGACCGT
20341 TCAACGGGGT GCTGCCGCGC TGGTGGGGTG TGCGGCCGAG GGCGATCACC TCCTCGACGG
20401 TGAAGTCGAG GTCGACGGCG CCGTCCTGGG TCATCGCCGC GATGAGCTGG GCGCTGCGGC
20461 GCATGGTCAG CGACGAGAGC TCCTGGCCGT CCACCTTCAC GGTGCCGGAG CTGGGTTTCA
20521 GGGCCCGGTA CACGCACCGC AGGGCGGTGG ACTTGCCGCT GCCGTTGGGG CCGACGAGGC
20581 CGACCACCTG ACCGCTGCCG ACGTCCAGGG AGAGGTCCCG TACCAGGCTC TTGCCGTCGG
20641 TCACCACCGA GAGCCCGTCG AGTTCGAGGT CCATCTCAAC GGCCTCCGAA CATGTAGGAC
20701 TTGCGGCGCA TCAGGGTGAT GAACACCGGG ACGCCGACCA GCGCGGTGAT GACGCCGAGC
20761 GGCAGCTCGC GGGGGGCGAC CAGGGTCCGC GACACGAGAT CGACCCAGAC CATGAAGACC
20821 GCCCCGGCGA GTGGTGCGAC GGCGAGCACC CGCGCGTGCG TCGCGCCCAC CACCATGCGT
20881 ACGAGGTGCG GCATGACGAG GCCGACGAAG GCGATGGAAC CGCTGACGGC GACCATCACG
20941 CCCGTCACCA GGGAGACGAG CACGAGCAGG GACTTGCGGT GTCGGTCGGG GCTGATGCCC
21001 AGGCTGGCTG CGGTCTCGTC ACCGAGAGCC AGGACGTCGA GCGGGCGGCC GTGCCGGTGC
21061 AGGACGAGGA CACCGAGCAG CACGGCGGCG GTGACCACCG GCAGCGAACC CCAGGAAGCG
21121 GCGCCGAAGC TGCCCATGGT CCAGTACAGG ACCATGCTGG TCGCCTCGGA GCTGGGCGCG
21181 AAGTAGATGA TGACACTCAT CACGGCCTGG AAACCCAGCG ACATGGCGAC ACCGGTCAGT
21241 ACGAGCCGCA GCGGCGAGAG CGCCCCCTTG GTGGACGAGG CGCCGTACAC CAGGACTGAG
21301 GCCACGAGCG CGCCGAGGAA GGCGCCCACG GACACCGCGT AGATCCCGAA CACGGCGAGC
21361 CCGCCCATGA CCGTCACACC GACGGCGCCC ACGGAGGCCC CCGAGGAGAC GCCCAGAACG
21421 AACGGGTCGG CCAGCGCGTT GCGCACCAGG GCCTGGATGG CGACACCGAC CGCGCTGAGC
21481 CCGGCCCCCA CGAGCGCCGC GAGCAGGACG CGCGGGGTGC GGATCTGCCA GATGATCTGG
21541 TACGTCGTCA CCTCGTCCGC CGAGATCGGC CCGCCACTGA GCGCGGCCCA GAGGAAGCGC
21601 GCGGTCTCGG CCGGGGGGAC CACGGCAGGC CCGAGACCGA TGGCGACGAC GACGGAGACG
21661 ACGAGCGCGG CGAACAGGCT CACGCAGATC GCCACCAGGC CCGTCCGGGA GCCGGTCCGA
21721 ACCGGCTCTT GCGCGGTGGG CGCGGGACGT TGCAGCGCCT CGGGTGGCGC GGGCGGTGAC
21781 ATGTGGATCG GCCTTCCGGT TTCGGAGCGT TGATGAACGG TGGATGTGCG TCCGTGGGGT
21841 GCCCGCGACC TTGGGCGGGC GCCCCGTCGG CTTCGGCTAC GCCGAACCGG GGATCTCGTC
21901 CTCGGAGCGC AGCACCAGGA GCCCGGCCAC CACGGCCACG GCGACGAGCA GCCCGAACGC
21961 GGCGGCGATC CCCGGGTACC CCGCGAGCCC GAGTCCCGCT CCGCCGAGGG CGGCGCCGGC
22021 GAAGACGCCG AGGCTCTGGC CCGCCGCGTT GAGGCTCAGC GCGGAACCCC GCATCGATCC
22081 GCAGCGCCTG ACCAGCAGAC TGACGGCGCA GGCGGCGACG GCCGCGTGGC TAGCGGCGTG
22141 CAGCGAAGTA AAGGCCAGGG CGAGCGGCAG CCAGGTCGTG AACCAGAAAC CGGTAGCGGT
22201 GACCAGGGCC GCCAACAGTC CGACGAGCAA GAGCTGTTCG GTACCCACGG TGGATTTCTC
22261 GGCGTTGGTG ATGCGGCCCG TGAGCAGGTT GCTGACGAAG AACGAGGCGC CGCTGAGCGT
22321 CCACACCAGC GAGAACAGGG CGGGGTCGAG GTGGAACCGG TCGTCGTAGT AGACCGCGAG
22381 GTAGGCGAGG TAGCCCATGA AGACCGCGGT GCGCAGGAAG GAGATGGCGA GCAGCGGCAC
22441 CGAGCCGCGG ACCTGGGCCA GGGCCTTGAA CGAGGCGAAG TAGCCCGTGC GCGGGCCACC
22501 CTCGACCACC GGGTCCTCGC CCTTCCTGCC GCGTACGAGG AAGACCGCGG CGAGCAGCAG
22561 CGAGACGACG GTGACGGCGA GCAGGTCGCC CTCCCATCCC CACAGCAGGG CCGGCAGGGC
22621 GATCAGGGGC GCGGCGAGCA TCGCCGTCAT CGAGGTCGTC GACGTGACGA GGGTGGCCGC
22681 ACGGGCGGCG GACTTGCCGT CGCCGAACCG GTCGGCGGCG GCAGCGGTGA GCGCCGGGTT
22741 GATCACCGCG GTGCCGGCGC CGACCAGCAG GCAGAACACC GCGGTCAGGA GGAAGTCTCC
22801 GCTCGCGCCG AGGGCTGAGG AGACGGCGAG TACGACGAGA CCGACCGCGA CCGCCTTCGA
22861 CTTGGGTACC CGGTCGATCA GGGGGGCCAG GGCCGTGCCC ACGGCGAGCG CCGCGAGGCC
```

```
22921 CCCCAGGCCG CGCAGGCCGC CCACCGCGGC GACACCGCTC CCGGTCTCCT CGGCGATCGG
22981 CACCAGATAC GTGCTGAAGA CGGTGAACGG CAGCAGGCCG ACGGCGGAGG CCACCAGGAC
23041 CGGCCACAGG GCTCGCGCCA TCTTCAGGTC GCCGGGCATC TCGGGGGACT TCTCCGGTGC
23101 GACGGCCGAA CGGGAGGTGC CGGCGCTCAC AGGTCACCGC CTGCGCGGTA GCGGTACATC
23161 GTCGTCTCGT CGGCGCTGAA CTGTGAGAAC GGGAAGGGCT CGGCGTTCAG GGCGGTGACG
23221 CCCGAGCCGA GGAACGCGCG GGCGACGGCG CTGCCCGTCT GGGCGTACAC GACGAGCGGC
23281 ACCTTCTGCT CGCGGCAGTG TTCCAGGATC AGGTCGAAGG TGCCGTTGGA GAGTGTCATC
23341 CCCGTGGCGA CGACGGCGTG GGCCTCTGCG AGGACCTCGG TCATGTCGTC CGCGACCGGC
23401 TCTCCCCACT GGGTGGTTCG CAGGTTGAGG TCGCACGGCA GGCAGACGCC GCCCCGCTCG
23461 CGGATCGCGG CGACGAGCGG GTTGACGACG CCGATGAGCG CGACCTTGGC GCCCTCCTCG
23521 ATGTCGAGCA GCCCGGCGAT GGACGCGTCC CGCGCCTTCG CCCGCACCTC GGGGGTCCCC
23581 ACCGGCAGCG GGACGGCCTC CTGCTCCGGG GCTTCCCGAT GCGGCTGTAT CTGTGCGAGG
23641 TAGGCGTCGA GCGCCGCTAT GCGCACCGGG GCGGACTCGT GGCGCAGCAA CTTCTCCAGC
23701 GGGTGCCCGG AGGCGTTCTC GCAGAAGTCC GGGGTGAGTT CGCCTGCCTC GAAGGAGCAG
23761 CCGCCGAAGG ACCGGCCGAC ACGCAGCACC AGGTAGTGGT TGTGGTACGT CACCGGTCCG
23821 CCGGCGAGCC GTGTCGTGTG GTAGAGCCAG AACGCGCTGG TGACGGTCAT GTCCTTCGGG
23881 TCGGGGCCGT AGTCCCCGGC GAGGACGGCA TCGGTGAGCT CGGCGACCGA CTGCGGCGTG
23941 GGAAAGGGCA TGTCAGAGGG CTTTCTTCTG GTCGGAGGTG GAGTCATCGG TCCACGTCAT
24001 GGCGGACCAG GGGTGGCTGA GCTCGTCGAG CGAGGTGATC TCGCGCGGTT CGAGGTCCTC
24061 GATGTCGGGC GCCTCGGTGT GCTTGGCGTA CGCGCTGTCG ACGTAGCGGT GACCTGTGTC
24121 CGCCGCGATG AAGACGTACG TCCGGGAATC GTCCTTCGAC CGCTCCCACC GGGTGGTCAG
24181 GTAGGCGGCG CCCGCGGACA GGCCTGCGAA GATGCCGCTG GAGCGGAGCA GGTGGACGGC
24241 GCCTGCGAGC GCGGAGTCGA AGCTGACCCA GTGGATCCGG TCGTACAGAT CGTGCCGGAC
24301 GTTCTCGAAC GGGATGGCGC TGCCGATGCC GGCGATGATC ATGTCCGGGT CCGAGACGTG
24361 CTCCGAGCCG AACGTGACGC TGCCGAAGGG CTGGACTCCG ACGAGGGAGA CGTCTCGGCC
24421 CGCCTCGCGC AGATACGAGG CGATGGCGCC TGTCGACGCG CCGGAACCCA CGCCGCCCAC
24481 CAAGGTCAGG GGCCCGGCGG GCACCTCGTC GGCGATCGTT TCGGCCACTT CGCGGTAGCC
24541 GTAGTAGTGG ATGCTGTCGT GGTACTGCCG CATCCAGTGG TACGAGGGGT TCTCCTCCAG
24601 GATCTCGGCG ATGCGCCGCA CCCGGAGCTC CTGGTCGAGG CGGAGATTCC TGGACGGCCG
24661 CACCTGCTCG AGCGTGGCAC CGAGAATCTC GAGCTGCGCC TTGAGCGTGC GGTCCACCGT
24721 GGTCGACCCC ACGATGTGGC ACTTCATGCC GTAGCGGTGG CAGGCGAGGG CGAGGGCCTG
24781 CGCGTAGATG CCGCTCGAAC TGTCGACGAG GGTGTCACCG GGTTTGACGG TGCCCGACTC
24841 AAGGAGGTGC CGCACCGCCC CCAGAGCCGA GTAGATCTTC ATGGTCTCGA ACCGCAGACA
24901 GACCAGGTCC GGCCGCAGTG CTATGAGATC GGGTTTCTTG ATCGCTTCAG CTATGTGCTC
24961 GTACATCTCC GTCTTCCGGT CGAGCGGGAC ATGAACCGTC TGCCTCGATC AGGTCCGGCT
25021 GGGCTGGGCC GCGGTGTGGC CGTGAGCCCG GACGAGAGCA TTATGGAAAT GAAAACGATT
25081 GTCAAAACCG AGTAAGGTGT GCGCCAGTCA TCACCACGGG AGCCGCACAG GCAGCTCTAC
25141 GCCCCGTGAC GGGCAGCAAG GCTTTTGGAG GAACTCATGC ATCTGCCCCG GGTCGGTCCG
25201 CGATCCTGCC TGTCGGGTCG GGCGGGCATG GACACTGGAG TGGGCACCGC CTACGGAACG
25261 TTCGGGGAAC TGCTCCAGGG TGAACTGCCG GAGGAGGCAG GCGATTTCCT CGTCACGCTG
25321 CCTGTCGCCC GGTGGGCGAG GGCGTCCTTC CGGTGCGACC CGGCCATGGG AGATGTCATC
25381 GTCAGGCCGT CGCACAAGGA GAAGGCGAGG CGGCTGGCCT GCCTGATCCT GGAGGAGGCA
25441 CCGGGGATGA CCGGTGGGGT GCTGACGGTC AACAGCGTGA TCCCGGAGGG CAAAGGGCTG
25501 GCCAGTTCAT CCGCCGACCT GGTCGCCACG GCGCGCGCGG TGGGGCGGGC CCTGCGGCTC
25561 GACATGCCGC CATCGCGGAT CGAGGGGCTG CTGAGGCTGA TCGAACCGAC CGATGGTGTC
25621 CTGTACCCGG GAATAGTCGC CTTCCATCAT CGAGCGGTGC GACTGCGCGC GATGCTGGGC
25681 TCGTTGCCCG CCATGTCGGT CGTCGGTGTC GACGAGGGCG GGCCGTGGA CACGGTCGAC
25741 TTCAACCGCA TACCCAAGCC GTTCACGCCG GCGGACCGGC GTGAGTACGC CGACCTGCTG
25801 AACCGGCTGA GTGGGGCCGT TCGCTCACGC GACCTCGCGG AGGTGGGCAG GGTGGCGACG
25861 CGCAGCGCGC TCATGAACCA GCCGCTTCGG TACAAGCGAC TGCTGGAGCC CATGCGGGAG
25921 ATCTGCAGGG ATGCCGGTGG TCTGGGCGTG GCCGTGGGCC ACAGTGGGAC GGCGCTCGGC
25981 GTGCTCCTGG ACGCCGCGGA TCCCGCGTAC CCGCACCGGG CCACCGCGGT GGCCCGGGCG
26041 TGCGGGGATC TGGCCGGGGC CGTCGCGGTC TATCGGACCC TCAGTTTCCC GAACGCCGTC
26101 AGCCATGGTG GTCGGACCGT CGGCTGAGGG CGGTTCCCGG AGGCATGCCC CGACGGGGCC
26161 CGATGGCGCG GCAAGCAGGG ATTCGCCTGA CGTTGAGGGT GGCCCGGATC GCTGTATGGT
26221 CACCGCGGTG CCGGTGCGTG GACCGTGTCA CTCCCGGCTC CCTTGTGAAG CCGATCGCCG
26281 GTGCTCCGCG GACGCTGTGA AGGTGGACGG CCTCGACCGG TTCGTCCAAG GGCCCGAGGT
26341 GCCAAGGCCT CTGCGACCGG TATCGCGGAC GCCCTCGGGC ACGTGGACTT CCTCTCGGCC
26401 GCCGCCGGGC CAACCGTTCC GGACAATCGA AGGGACCCAG GTTCATGCTC ACCGCACAGC
26461 AGCCTGCTCC CGGCGTCGTG CCCGCCCGGA TCCACGTCAC GGACAGGTTG GAGGCCGCTC
26521 ACCCGCTCGC CGCTGACGGG GCTGTCGTCC TGACAGGCGT CGAGCCCTCC GGTGACGGCC
26581 TGGTCCTCGC CGCCGCAGCC GTCCTGGGGG AGCGGCTGCA GCAGGTGTTC CCTCACCGGC
26641 TGCGGGCGTC CGACGGCTCG AACTTCGTCC ACCTTCATGC GGACAGCTTC GACTTCGTCG
26701 TCAACGTAGG GGGCGTCGAG CATCGCCGAC GTGATCCGGA TGAGGACTAT GTCCTCATCC
26761 AGTGCGTCCG GCAGTCCGAC TCCGGCGGCG ACTCCTTCGT GGCTGACGCC TATCGCTTCG
```

```
26821 TGGACCACTG CGCGACGGCC GATCCTGAAC TGTGGGACTT CCTGACCCGA GGGGACGTCG
26881 ACCTGTACGG CGCGTGGTCC GGACTGCGTG GTATGCCCGC AACCCCCTTT GTGGGCAGGC
26941 ATGTCGAGTA CACCCGCGCC GGTCGGCGTA TCGTCCGGCG CGGCGACGGG GTGACCCCTC
27001 TGCACCGGGA CCCTGGCGCG GACCACACCC GGCGGATGCT CGCCCGTCTG GAGGAAGCCG
27061 TCCATGCGCT GGAGGAGACG CTCCCGCGAT TCCGGCTCGA CAAGGGCGAA ATCCTCGTCC
27121 TGGACAACTA CCGCTGCTGG CACGGCCGCG AGGCTCACAC GGGAGATCGC GCGGTACGTA
27181 TCCTCACGGT GCGCAGCAGC GACGCCCGCT GAGGCGCTGT TGGTTCGCCT CACTCGCCGT
27241 GACACAGGGG CAGGCGTCTG CGGCGGTGCT GTTTCCGCGC GGGACGGACC GGGGGAGATT
27301 CCCCGGTCGG TAAAGGGGGC GACCGGCGAT CCGCTCACCC CGCCTCGATC ATTGCGCAGG
27361 CTCTTCGAGC GCTTCGTGCT TCACGCCGGC TGCCAGATCC GGGCCAGTGC CTCCGGGGTG
27421 AGTACTTCCT CCGGTGATCC CTGCCCGATC AGTCGTCCGT CGGCCAGGAG CAGGCAGGCG
27481 TCGGCCGAGC GGGCGGCGTC CAGGTCGTGG GTGGCCTGGA CGACGGTGGT GCCGTCGGCG
27541 ACCAGGTCCG TCAGCAGGGC CGTGATCCGC TCCCGCGCCT CGGGGTCGAG TCCGGTGGTC
27601 GGCTCGTCCA GGAGAAGCAG GTCGGACTGT TGGGCGAGGC CCTGCGCGAT CAGCACGCGC
27661 TGACGCTGGC CGCCCGACAG CTCGCCGAGC TGGCGGGCGC CGAGGTCGGC GACCCCCAGC
27721 CTCTCCATGG CGGAGTCGAC CGCGGTCCGG TCCGTGCGGG TCAGCCGCCG CCACAGGCCC
27781 CGCTGTCCCC AGCGGCCCAT CTCCACCGTC TGCCGCGCCG TGAGGGGGAG GGTGTCGCCG
27841 ACGGCACCGC GCTGCGGGAC GAAAGCCGGC GGGGAGCCCT CTGCGTACCG GAGTTGTCCG
27901 GATGTGGCGG TGATCACTCC GGCCAGGACG CCCAGCAGCG TCGACTTGCC GCTTCCGTTG
27961 GGTCCGACCA GGGCGGTCAT GGCCAACGGC GGTATTGCGG CGCTGAGTTG GTGGAGCACG
28021 GGGCGGCCGG GGTAGCCGGC GCTCAGCCGC TGGAACCGGA CGCGTTCATT CCGCAGTTCG
28081 GTGGCCGGCG GGAACGGAGG GTTGTTATTG AACATGGTTG TCATTATATG GTCCTCGTAT
28141 GGAGTGGTTG ACGGCCCCTT TCGAGGTGGC CTTTGTGCAG AGGGCCCTAT GGGCCGGGAT
28201 CCTGGTGTCG GCGATATGCG CCCTCGCGGG AACGTGGGTG GTGCTGCGCG GGATGGCCTT
28261 CCTCGGTGAC GCGATGTCGC ACGGGCTGCT GCCCGGCGTC GCGGTCGCCT CCCTGCTGGG
28321 AGGCAACCTG CTGGTGGGGG CGGTGGTGAG CGCGGCCGTG ATGGCGGCGG GCGTCACGGC
28381 CCTCGGGCGG ACTCCGCGAC TGTCCCAGGA CACCGGCATC GGCCTGCTGT TCGTGGGCAT
28441 GCTGTCGCTC GGCGTCATCA TCGTGTCGCG GTCGCAGTCC TTCGCGGTGG ACCTCACCGG
28501 CTTCCTGTTC GGAGACGTCC TCGCCGTGCG GGGGAGCGAT CTGCTGCTTC TTGGAGTAGC
28561 CCTGCTGCTG GCGCTGGCCG TCTCGGTGCT CGGCTACCGG GCTTTCCTGG CCCTCGCGTT
28621 CGACGAGCGC AAGGCCCGGA CACTCGGGCT GCGTCCCCGG CTCGCCCATG CCGTGCTGCT
28681 CGGCCTGCTG GCGCTGGCCA TCGTGGCCTC CTTCCACATC GTGGGCACGC TGCTCGTCCT
28741 CGGTCTGCTC ATCGCCCCGC CCGCGGCGGC CATGCCCTGG GCGCGAAGCG TCCAGGCGGT
28801 CATGGTCCTC GCGGCGCTCC TCGGCGCCGC CGCCACCTTC GGCGGCCTGC TCCTGTCCTG
28861 GCATCTGCGC ACCGCGGCCG GAGCGACCGT CTCGGCCCTC GCCGTCGCTC TCTTCTTCCT
28921 GTCCCACCTG GCATCCGGAC TTCGGCACCG CCGCCGTGCG CGCCGGGGCG GTCTTGCCGA
28981 ACCGGCGGTC GCCCCGGGCC GCGACCTCCT CCACGTCCTG ACCGAGAGAA ACCTGAGGCG
29041 ATCTCCTTGC TCGTCCGAAA AAACGTCACA TCGCTGGCTC CGGCGCTTGC GGCCGTGATC
29101 CTCCTGACCG CCGGATGCGG GGGCGGGGAC GAGGCCAAGT CCGGTTCCGG GCCCGCCTCT
29161 TCGTCCCCCA CTCCGCACGG CTATGTCGAA GGCGCCACCG AGGCGGCCGA GCAGCAGTCC
29221 AGACTTCTGC TCGGCGACCC CGGGAGCGGT GAGACCCGCG TGCTGGACCT GATCACCGGC
29281 AAGGTGTACG ACATCGCCCG CAGCCCCGGT GCCACCGCAC TCACCACGGA CGGCCGCTTC
29341 GGCTACTTCC ACGGCCCGGA CGGCATACGG GTGCTCGACA GCGGTGCGTG GATGGTGGAC
29401 CACGGCGACC ACGTCCACTA TTACCGCGCG AAGATCAAGG AGGTCGGCGA ACTCCCGGGC
29461 GGCACCGGTA CGAGCATCCG CGGCGACGCG GGCGTGACCG TGGCCTCGTC GGCGGACGGG
29521 AAGGCGAGCG TGTATCGCAG GGCGGACCTG GAGAAAGGCG CCCTGGGCAC GCCGTCCCCG
29581 CTGCCCGGCA CGTTCGCCGG CGCCGTCGTG CCGTACGCGG AACACCTGGT GACACTCACC
29641 GCTGAGAGCG GGGCTCCGGC GAAGGTCGCC GTGCTGGACC GTTCCGGCAA GCGCGTCGCC
29701 GCTCCGGAGG CGGAGTGCGA GGAGCCTCAG GGCGACGCGG TCACCCGGCG CGGGGTTGTC
29761 CTCGGCTGCG CCGACGGCGC TCTGCTCGTC CATGAGGACG ACGGCGCCTT CACGGCGGAG
29821 AAGATTCCGT ACGGCGAGGA CGTGCCGAAG ACCGAGCGGG CCGTGGAGTT CCGGCACCGC
29881 CCGGGCAGCA GCACCCTCAC GGCACCCGCC GGCAAGGACG CTGTCTGGGT CCTGGATGCC
29941 GGCGAGGGCG CCTGGACCCG GGTGAAGACC GGCCCCGTGG TCGCCGCCAA CACGGCCGGC
30001 GAAGGCTCGC CGCTGGTCGT CCTGGAGACC GACGGGGCCC TGCACGGCTA CGACATACCC
30061 ACCGGCAAGG AGACCGGCGT GACCGATCCC CTGCTCAAGG AACTGCCCGG AACCGGTGCG
30121 GGCGGCGGCG CGGCTCCGGT GATCGAGGTG GACCGCAGCC GGGCCTACCT CAACGACCCC
30181 GAGGGCAAGC GCGTGTACGA GATCGACTAC AACGACGATC TCCGCGTGGC CCGTACGTTC
30241 GACGTCGACG TACGGCCGTC CCTGATGGTG GAGACGGGCC GATGAGCGCG CGCGTGGGCG
30301 CTCCACGGAT GCGTGCCCTG CTGGTGTCCC TGGCCGGATT CTTCGTCGTC GCCGGTGCGG
30361 CGACCGGCTG CGCGGGCGGC GGAGACGAAC GGCCCCGGGT CGTGGTGACC ACCAACATCC
30421 TCGGCGACAT CACCCGGGAG ATCGTCGGGG ACGAGGCCGG CGTCAGTGTC CTGATGAAGC
30481 CCAACGCCGA CCCGCACTCC TTCGGCCTCT CGGCCGTGCA GGCCGCTGAG TTGGAGAACG
30541 CCGACCTGGT CGTCTACAAC GGGCTCGGCC TGGAGGAGAA CGTGTTGCGG CACGTGGAGG
30601 CTGCCCGCGA GTCCGGAGTG GCCGCCTTCG CCGCGGGTGA GGCGGCCGAC CCGCTCACCT
30661 TCCATGCCGG ACAGGACGGC GGCCCCGAAG AGGACGCCGG CAAGCCCGAT CCGCACTTCT
```

```
30721 GGACCGACCC CGACCGCGTA CGCGAGGCCG CCGGCCTGAT CGCCGACCAG GTCGCCGAGC
30781 ATGTGGAGGG CGTCGACGAG AAGAAGGTCC GGGAGAACGC CGAGCGGTAC GACGGACAAC
30841 TCGCCGACCT CACGGGATGG ATGGAGAAGT CCTTCGCCGC CATCCCCGAG GACCGGCGTG
30901 CCCTGGTGAC CAACCACCAC GTCTTCGGCT ACCTCGCCGA CCGCTTCGGC CTCCGCGTCA
30961 TCGGCGCGGT CATCCCCAGC GGAACCACGC TCGCCTCGCC CAGCTCCTCC GACCTGCGCT
31021 CTCTCACCCA GGCCATGGAG AAGGCCAAGG TGCGCACCGT CTTCGCCGAC TCCTCCCAGC
31081 CCACCCGGCT CGCCGAGGTC CTGCGCCAGG AGATGGGCGG CGACGTGGAC GTCGTCTCGC
31141 TCTACTCCGA GTCGCTGACC GAGAAGGGCA AGGGCGCCGG AACCTACCTG GAGATGATGC
31201 GCGCCAACAC CTCCGCCATG GCCGAGGGCC TCACCGGCGA CTGAACGAGC TTCCCCGCGG
31261 CACGGCACTT CGAGCGCCGG CCGCTCCACC CCACAAACCC GCGCCTGAGG GCCGGAGAGG
31321 AAACACCGAT CATGAACAAG CCCACCCGCG CCAGAGTCTT CACGGGCACG GCGCTGGTCG
31381 TGGCGGCGTC GATGGCGCTG ACCGCCTGCG GCGGCAACGG CAACGACGAC GCCCCTTCCG
31441 GCAAAGAGCC CAAGGAGCAG AAGAGCAGCG AGGCCGCGGC GGTCGGGAAC CCGATCGTCG
31501 CCTCGTACGA CGGGGGACTG TACGTCCTCG ACGGCGAGAC CCTGAAGCTC GCGAAGACGA
31561 TCGCACTGCC CGGCTTCAAC CGGGTCAACC CGGCGGGCGA CAACGAGCAC GTCGTCGTCT
31621 CCACGGACTC CGGCTTCCGC GTGTTCGACG CCACCCGACA GGAGTTCACC GACGCCGAGT
31681 TCAAGGGTTC CAAGCCGGGG CACGTCGTCC GGCACGGCGG CAAGACGGTC CTGTTCACCG
31741 ACGGCACGGG AGAGGTGAAC GTCTTCGACC CCGCCGACCT GTCCGACGGG AAGAAGCCGG
31801 ACGGCCGCAC CTACACGTCC GCGAAGCCCC ACCACGGTGT CGCCATCGAA CTGGCCGGCG
31861 GAGAACTCGT CACCACCCTC GGCACCGAGG AGAAGCGCAC CGGAGCCCTC GTCCTGGACA
31921 AGGACAACAA GGAGATCGCA CGCGCCGAGA ACTGCCCCGG AGTGCACGGC GAGGCCGCCG
31981 CCCAGGGCGA GGTGGCCGGC TTCGGCTGCG AGGACGGCGT CCTGCTCTAC AAGGACGGCA
32041 AGTTCACCAA GGTCGACGCC CCCGGCGACT ACGCCCGCAC CGGCAACCAG GCCGGCAGCG
32101 ACGCCTCCCC GATCCTCCTC GGCGACTACA AGACCGACCC CGACGCCGAA CTGGAACGCC
32161 CCACCCGCAT ATCCCTGATC GACACCCGTA CGGCGAAGAT GAAGCTGGTC GACCTCGGCA
32221 CCAGCTACTC CTTCCGCTCC CTCGCCCGCG GCCCGCACGG CGAAGCCCTC GTGCTCGGCA
32281 CCAACGGCAC CCTCCACGTC ATCGACCCGG AGACCGGAAA GGTCGAGAAG AAGATCGACG
32341 CGGTCGGCGA CTGGACCGAG CCCCTGGACT GGCAGCAGCC CAGGCCCACC CTGTTCGTCC
32401 GGGACCACAC GGCGTACGTC TCCGAACCGG GCAAGCGCCA ACTCCACTCC ATCGACCTGG
32461 AATCGGGGAA GAAGCTGGCA TCCGTCACCC TGCCGAAGGG CACCAACGAA CTGTCCGGCA
32521 CGGTCGCCGG TCACTGACCT GTCCCGTTCC CTCTTTTCCT CGGGCCCCGA GGAGCGCAAC
32581 GCCTGCCGGA TTCGTGTTCC GGCAGGCGTT GCTGTCGTCG GAGCCTGCAA CCTTGACGAC
32641 CCTGCCGAGG AGAACCGTTT CACCACGGAG GCCTGGGGTG CGCAGATGGA ACTGTGCGCG
32701 CTCCACTCCA GGGACCGTGA CGCCACCGTC AAGACCTGTG CCGCCGGCCG CCCGAAACGC
32761 AAGCCGTCGT ACGGCTTCCT GGGCCGTCCC ACAGCCGCCG AGGAGCTCGC CGCGGTCACG
32821 AGCTGCGGCG GCGGTGCCTG CGCCGCCACC ACACGATCGC GAGCGTGAAG GCGGCCGCAA
32881 CGCCCAGCAG GGCCCACAGG ATGGTGGAGA GCACGCTCTC GGCCTCGCGC AGGGAGGTCG
32941 AGACCAGTGT TCCCGCGGAC ACGTAGAGCG CGGACCACAT CGCGGCTCCG GCGAGGGAGG
33001 CGGGCAGGAA GCGGAGGTAG CGCACGGAGC CGACGCCGGC GGTCGCGGGG GTGAGGGTGC
33061 GTACCACGGG CAAAAGGCGG GTCAGGAAGA CGGCGCGCGC CCCGTACCGG TGGCAGAGCT
33121 CTTGCGCGCG GTCCCAGTGG TGCTGCCCAA TCCGCCGTAC CAGGCGCGTC TCCCGCATCC
33181 GCTGCCCGTA GCGGATGCCG AGGAAGTAGC CGATGTGGTC GCCGGCCGAG CTGCTGAGTG
33241 TGACGACGAG GAAGAGGGCC AACAGCGGGC GTGTCCCCTC CGTTCCGGCG CTCAGGGCCA
33301 GTACCGCGAC CTCGCCGGGG ACGGCCATGC CGGCCCCAAG GCCGGATTCC GCGAACGCGA
33361 ATACGGAGGC CAGCGCGAAT CTGGTGACCG GGTTCATGTC CGACACCGCT GTCAGTACAT
33421 CGTTCATCCA CGACACGGCA GCCCCGCTCT GTCTCTCCTC GTTCGTGGAG CCCTCCCGAC
33481 GGCGCCACGG GGATTCCCGC GCCCTTCTTC CGAGAACACA CCGAAGAGAA CAGCGGAACG
33541 ACTTCCCGGC GTCACCGGAC GCATACCCGG GCGGCCGGTG GGAGCGCCTG AAAAAGAACG
33601 AAGGGACACC AACCTACCAG GGAACCGCTG GACGACTCCT CCCTCCCGGC CACGACCACC
33661 CCGCGACGGA CCCCGCAGAC CGCCCCCGGC AACCATTCCC CTTCACCCAC CCCGTCCGCC
33721 GACGGAGCAC GGGGGCTCGC CGTACAGATC CGGGCCTCGT TGATCCACTG GGTGAGAACG
33781 GCGGGGCCGG CCCCGGCCGC GAGGGCGGCC CGGTAGTGAG ACAGACGCTT CTCGCCCTTT
33841 CTCACCGCCC GCCGGGCCTG CTCGACCTCC GGGGCGCGGC CATCGGATGC GGCAGCCGCG
33901 TGCGTCAGGG CGGTGAGGGT GGCGGTCAGA CGTTCCGGCG CGAAGGCACG GGCGATCCAC
33961 TGGTCGAGTG CCGGGCAGAT CATGTCCTCC CGCAGGCACA TCATGTCCTC CCGCGGGCAC
34021 ACGGTGCGGG GGTGACCGAG TCCGGGGTGG AGGGCCTTGT TCCTGGGACC CGCTCCTGAC
34081 CGTGTACGGG CGTCCGAGGT CGGCTCAGGC GATCGCGGTC AACTACCCCG TGGGCTACAG
34141 TGCGTTGACT GCGGGCAGTG CACACGCCCA CCGGCACCGA CGACGCGGAG AAGCATGGGC
34201 GGGAGCGCGA TCAGGACCCG GCAGCTGACC AAGCACTTCG GTGCGGTGCA GGCGCTGGTC
34261 GGCGTGGATC TGGAGGTGCC CGCGGGGAGC GTGCTGGGGC TCCTGGGACA CAACGGTGCC
34321 GGGAAGACCA CGCTGATCCA GATCCTCTCG ACGGTGCTCC CCCCGTCCGG TGGGTCCGCC
34381 GAGGTCGCCG GCTTCGACAT CGTGCGCGAT GCCCGACGGG TACGCGCCTG TATCGGGGTG
34441 ACGGGGCAGT TCGCTGCCCT GGACGAGCAT CTGTCCGGGC TCGCCAATCT GGTGCTGATC
34501 TCCCGGCTGC TGGGTGCCCG GCCGAGGGAG GCCAGACGCC GGGCGGCCGA ACTGGTCGAA
34561 CAATTCGGTC TCACCGAGGC AGCGGACAGA CCGATGCGGA CCTACTCCGG CGGAATGCGG
```

```
34621 CGGCGCATCG ACCTGGCGGC GAGTCTGGTG GCCAGGCCCT CGGTGCTGTT CCTCGACGAG
34681 CCCACCACCG GGCTGGACCC GGTGAGCCGC ACCGCACTCT GGGAGACGGT GGAAGGGCTG
34741 GTCGCCGAGG GCACGACGGT TCTGCTGACC ACCCAGTACC TCGACGAGGC CGACCGGCTG
34801 GCGGACCGGA TAGCGGTGCT GTCGTCCGGC CACGTGGTGA CGGTCGGCAC GGCGGCGGAG
34861 CTCAAGGCGG CGGGCACCCG GTCCGTCCGC CTGACCTTCG GGTCCGCGGC GGATCTGGAG
34921 AGCGCGGAAG GAGCGCTGCG CCTGGAGGGC CTCGGCCTCA CAACGGATCC GGTGTCCCGG
34981 ACGGTGTCAC TGCCGCTGGC GGCAACGGCC GAGCTGGCCG GGATCTTCCG GATTCTCGGC
35041 GCGGCGGGCG TGGAGCTCGC CGAACTGGCG CTCAAGGAGC CCACGCTGGA CGACGTGTAT
35101 CTGAGCCTGG CGGAGAGCTG GGAGACCACG AGCGGGGGAA CGGTCCGGTG CTGACCACAC
35161 GACGTACGGG TCCGGGGACC TCGCCGGTGG CGGACGGGCC CGGGTGGCGC GGCGGGGGTG
35221 CGGGGATCGG CACCCAGTTC CGGGTGCTGA CCGGCCGGCA GTTCCGGATC ATCTACGGGG
35281 ACCGGCGGAT CGCGCTGTTC AGCCTGCTCC AGCCGATCAT CATGCTCATG CTGTTCAGTC
35341 AGGTGCTGGG CCGCATGGCC AATCCGGAGA TCTTCCCGCC GGGTGTGCGC TACCTCGACT
35401 ACCTGGTGCC GGCTCTGCTG CTGACGACCG GGATCGGTTC CGCGCAGGGC GGCGGGCTGG
35461 GTCTCGTCAG GGACATGGAG TCCGGGATGA TGGTCCGGCT GCGGGTGATG CCGGTACGGC
35521 TGCCGCTGGT CCTGGTGGCC CGGTCGCTGG CCGATCTGGC GCGGGTCGCC CTGCAGCTCG
35581 TGGCGTTGCT CGCCTGTGCG ATGGGGCCGC TGGGCTACCG GCCGGCCGGG GGCGTGTCGG
35641 GGATCGTCGG CGCGACGCTG CTCGCGTTGC TCGTCGCGTG GTCGCTGATC TGGGTGTTCC
35701 TGGCCCTCGC CGCGTGGCTG CGGAGCATCG AGGTGCTGTC CAGCATCGGG TTCCTCGTCA
35761 CCTTCCCCCT GATGTTCGCG TCGAGTGCCT TCGTCCCGCT CGACATTCTG CCGGGATGGC
35821 TCAGGGTCAT CGCGACGGTC AATCCCCTCA CGTACGCGGT GGAGGCGTCC CGCGATCTGG
35881 CGCTGGACCA CAGCGCGCTG GGCGCGGCGC TCGCGGCCGT CGGCACCAGT CTTGCGCTCT
35941 TGGCGGTGAC CGGTCTGCTG GCGGTACGCG GGCTGCGGCG CCCGCCGGGT GCGGGCGGCC
36001 CGCACCGGAC GCCCTGACCC CTCCCCACCA CCTGCCCAGT GTGACGTTTG CGCAGATGAG
36061 AACGTGCGTA AACGCCGCAT ACGCAAAGAT CGTCCCTGCC GGGACCCATT GACGTTCGCA
36121 GGGGCGTGGA ACATACTGGC GATCAAGTCG CACAGGAACC AACAGGCACA CCAACCACAG
36181 GCGTTACAGG GGGGGTTGGT GTTTCGTCCA TATCAAGTGG TTTGGTCCGC CGAAGCGGTT
36241 GGACCTCACA TGACGGCAAC AGGGCATTCG CACATGCCTG ATGACGGGAC GGCACACCTC
36301 ACGCAGCGGC GACCGGTCGC AAGCCGGACG CGGAATGACT CCCTGCCTTA CAGGTATGCG
36361 AGCGCGGATG CGTCGTTCGA CCGGAGTCAG GAGGGGGAGT GCCTGCCGTG AGTGAGAGCC
36421 GCTGTGCCGG GCAGGGCCTG GTGGGGGCAC TGCGGACCTG GGCACGGACA CGTGCCCGGG
36481 AGACTGCCGT GGTTCTCGTA CGGGACACCG GAACCACCGA CGACACGGCG TCGGTGGACT
36541 ACGGACAGCT GGACGAGTGG GCCAGAAGCA TCGCGGTGAC CCTCCGACAG CAACTCGCGC
36601 CGGGGGGACG GGCACTTCTG CTGCTGCCGT CCGGCCCGGA GTTCACGGCC GCGTACCTCG
36661 GCTGCCTGTA CGCGGGTCTG GCCGCCGTAC CGGCGCCGCT GCCCGGGGGG CGCCACTTCG
36721 AACGCCGCCG TGTCGCGGCC ATCGCCGCCG ACAGCGGAGC CGGCGTGGTG CTGACCGTCG
36781 CGGGTGAGAC CGCCTCCGTC CACGACTGGC TGACCGAGAC CACGGCCCCG GCTACTCGCG
36841 TCGTGGCCGT GGACGACCGG GCGGCGCTCG GCGACCCGGC GCAGTGGGAC GACCCGGGCG
36901 TCGCGCCCGA CGACGTGGCT CTCATCCAGT ACACCTCGGG CTCGACCGGC AACCCCAAGG
36961 GCGTGGTCGT GACCCACGCC AACCTGCTGG CGAACGCGCG GAATCTCGCC GAGGCCTGCG
37021 AGCTGACCGC CGCCACTCCC ATGGGCGGCT GGCTGCCCAT GTACCACGAC ATGGGGCTCC
37081 TGGGCACGCT GACACCGGCC CTGTACCTCG GCACCACGTG CGTGCTGATG AGCTCCACGG
37141 CATTCATCAA ACGGCCGCAC CTGTGGCTAC GGACCATCGA CCGGTTCGGC CTGGTCTGGT
37201 CGTCGGCTCC CGACTTCGCG TACGACATGT GTCTGAAGCG CGTCACCGAC GAGCAGATCG
37261 CCGGGCTGGA CCTGTCCCGC TGGCGGTGGG CCGGCAACGG CGCGGAGCCC ATCCGGGCAG
37321 CCACCGTACG GGCCTTCGGC GAACGGTTCG CCCGGTACGG CCTGCGCCCC GAGGCGCTCA
37381 CCGCCGGCTA CGGGCTGGCC GAGGCCACCC TGTTCGTGTC GAGGTCGCAG GGGCTGCACA
37441 CGGCACGAGT CGCCACCGCC GCCCTCGAAC GCCACGAATT CCGCCTCGCC GTACCCGGCG
37501 AGGCAGCCCG GGAGATCGTC AGCTGCGGTC CCGTCGGCCA CTTCCGCGCC CGCATCGTCG
37561 AACCCGGCGG GCACCGTGTT CTGCCGCCCG GCCAGGTCGG CGAGCTGGTC CTCCAGGGAG
37621 CCGCCGTCTG CGCCGGCTAC TGGCAGGCCA AGGAGGAGAC CGAGCAGACC TTCGGCCTCA
37681 CCCTCGACGG CGAGGACGGT CACTGGCTGC GCACCGGCGA TCTCGCCGCC CTGCACGAAG
37741 GGAATCTCCA CATCACCGGC CGCTGCAAAG AGGCCCTGGT GATACGAGGA CGCAATCTGT
37801 ACCCGCAGGA CATCGAGCAC GAACTCCGCC TGCAACACCC GGAACTTGAG AGCGTCGGCG
37861 CCGCGTTCAC CGTCCCGGCG GCACCTGGCA CGCCGGGCTT GATGGTGGTC CACGAAGTCC
37921 GCACCCCGGT CCCCGCCGAC GACCACCCGG CCCTGGTCAG CGCCCTGCGG GGGACGATCA
37981 ACCGCGAATT CGGACTCGAC GCCCAGGGCA TCGCCCTGGT GAGCCGCGGC ACCGTACTGC
38041 GTACCACCAG CGGCAAGGTC CGCCGGGGCG CCATGCGTGA CCTCTGCCTC CGCGGGGAGC
38101 TGAACATCGT CCACGCGGAC AAGGGCTGGC ACGCCATCGC CGGCACGGCC GGAGAGGACA
38161 TCGCCCCCAC TGACCACGCT CCACATCCGC ACCCCGCGTA ATCGCCGGAG GGCGGCCCTG
38221 CCCTGGAACG GGCACCGCGG TGCCCGCCGA CAGCGAGGAG TAGCTCCACA TGAACCCGCC
38281 CGAAGCGGTC AGCACGCCCA GCGAGGTCAC CGCGTGGATC ACCGGACAGA TCGCCGAGTT
38341 CGTGAACGAG ACACCCGACC GGATCGCCGG TGACGCACCC CTGACCGACC ATGGCCTCGA
38401 CTCCGTCTCC GGAGTTGCCC TCTGCGCGCA GGTCGAGGAC CGCTACGGGA TCGAGGTCGA
38461 CCCGGAGCTG CTGTGGAGCG TCCCCACACT CAACGAGTTC GTCCAGGCAC TGATGCCCCA
```

```
38521 GTTGGCCGAC CGCACCTGAG GGGATCCGCG AGAGATGGAC ATGCAGTCGC AGCGCCTCGG
38581 CGTCACCGCC GCCCAACAGA GCGTCTGGCT CGCCGGCCAG CTGGCGGACG ACCACCGCCT
38641 GTACCACTGT GCGGCGTACC TGTCACTCAC CGGGTCCATC GACCCGCGGA CACTCGGCAC
38701 GGCGGTCCGG CGGACCCTCG ACGAGACCGA GGCGCTGCGT ACCCGGTTCG TACCGCAGGA
38761 CGGGGAACTG CTGCAGATCC TCGAACCCGG TGCCGGACAG CTCCTGCTGG AAGCCGACTT
38821 CTCCGGCGAC CCGGACCCCG AGCGGGCGGC ACACGACTGG ATGCACGCGG CGCTCGCCGC
38881 ACCGGTCCGC CTCGACCGCG CCGGGACCGC CACCCACGCC CTGCTCACCC TCGGCCCGTC
38941 CCGCCACCTG CTGTACTTCG GCTACCACCA CATCGCGCTC GACGGCTACG GTGCCCTGCT
39001 CCACCTGCGC CGCCTCGCCC ACGTCTACAC CGCCCTCAGC AACGGGGACG ACCCCGGCCC
39061 CTGCCCGTTC GGCCCCCTGG CCGGTGTCCT CACGGAGGAG GCGGCCTACC GTGACTCCGA
39121 CAACCATCGG CGCGACGGGG AATTCTGGAC CCGGTCCCTC GCCGGTGCGG ACGAGGCCCC
39181 CGGGCTGAGC GAGCGGGAGG CCGGCGCTCT CGCCGTCCCG CTGCGCCGCA CCGTGGAGCT
39241 GTCCGGCGAA CGGACGGAGA AGCTGGCCGC CTCGGCCGCG GCCACTGGAG CTCGCTGGTC
39301 GTCACTGCTC GTCGCCGCCA CCGCCGCGTT CGTACGCCGC CACGCTGCCG CCGACGACAC
39361 CGTCATCGGC CTGCCCGTCA CCGCCCGGCT CACCGGGCCG GCGCTGCGTA CCCCGTGCAT
39421 GCTCGCCAAC GACGTGCCGC TGCGCCTCGA CGCCCGGCTC GATGCCCCGT TCGCCGCGCT
39481 CCTTGCCGAC ACCACCCGCG CCGTCGGCAC GCTGGCGCGC CACCAGCGGT TCCGCGGGGA
39541 AGAACTCCAC CGGAACCTGG GGGGCGTCGG CCGCACCGCG GGCCTGGCGC GGGTCACCGT
39601 CAACGTCCTG GCGTATGTCG ACAACATCCG GTTCGGCGAC TGCCGGGCCG TGGTCCACGA
39661 GTTGTCCTCG GGACCGGTCC GCGACTTCCA CATCAACTCC TACGGCACCC CCGGCACCCC
39721 CGACGGCGTC CAGCTGGTCT TCAGCGGTAA CCCCGCCCTG TACACGGCCA CCGATCTGGC
39781 CGACCACCAG GAGCGGTTCC TGCGCTTCCT CGACGCTGTG ACCGCCGACC CGGACCTGCC
39841 GACCGGAAGA CACCGCCTCC TGTCGCCGGG CACCCGCGCC CGGCTGCTCG ACGACTCCCG
39901 CGGCACGGAA CGCCCCGTAC CGCGTGCCAC CTTGCCGGAA CTCTTCGCCG AACAGGCCCG
39961 GCGCACCCCC GACGCGCCCG CCGTCCAGCA CGACGGCACC GTCCTCACCT ACCGCGACCT
40021 GCACCGGAGT GTCGAACGGG CGGCCGGACG GCTGGCCGGC CTCGGCCTGC GTACCGAGGA
40081 CGTCGTCGCC CTCGCCCTCC CCAAGTCCGC CGAGAGCGTC GCGATCCTGC TCGGCATCCA
40141 GCGGGCCGGC GCCGCCTACG TGCCGCTGGA CCCCACCCAT CCGGCCGAGC GGCTGGCCCG
40201 TGTACTCGAC GACACCCGAC CCCGGTACCT CGTCACCACC GGACACATCG ACGGCCTGTC
40261 CCACCCCACG CCGCAGTTGG CCGCCGCCGA CCTCCTCCGT GAGGGCGGCC CAGAGCCCGC
40321 CCCGGGCCGC CCGGCACCCG GCAACGCGGC GTACATCATC CAGACCTCCG GCTCCACCGG
40381 ACGGCCGAAG GGTGTCGTCG TCACTCACGA AGGGCTGGCC ACCCTCGCCG CCGACCAGAT
40441 CCGGCGCTAC CGCACGGGAC CGGACGCCCG CGTACTGCAG TTCATCTCCC CGGGGTTCGA
40501 CGTCTTCGTC TCCGAACTGA GCATGACCCT CCTGTCCGGC GGCTGCCTGG TGATACCGCC
40561 GGACGGCCTG ACCGGCCGTC ACCTCGCCGA CTTCCTTGCC GCGGAGGCCG TCACCACCAC
40621 ATCCCTCACC CCCGGCGCAC TCGCCACCAT GCCCGCCACA GATCTCCCGC ACCTGCGGAC
40681 TCTGATCGTC GGCGGAGAGG TCTGCCCGCC GGAGATCTTC GACCAGTGGG GCCGGGGCCG
40741 GGACATCGTC AACGCGTACG GGCCCACCGA GACAACCGTC GAGGCGACCG CCTGGCACCG
40801 TGACGGTGCC ACCCACGGCC CCGTCCCGCT CGGCCGCCCC ACCCTCAACC GGCGCGGCTA
40861 CGTCCTCGAC CCGGCGCTCG AACCCGTCCC CGACGGGACG ACCGGCGAAC TGTACCTGGC
40921 CGGCGAGGGC CTCGCCCGGG GCTACGTCGC TGCTCCCGGG CCCACCGCCG AGCGTTTCGT
40981 CGCCGACCCG TTCGGCCCGC CCGGCAGCCG CATGTACCGC ACCGGTGACC TGGTGCGGCG
41041 GCGCTCCGGC GGCATGCTGG AATTCGTCGG ACGAGCCGAC GGACAGGTCA AACTCCGCGG
41101 CTTCCGCATC GAACTCGGCG AGGTCCAGGC CGCGCTCACC GCTCTCCCCG GGGTACGTCA
41161 GGCCGGCGTC CTGATCCGCG AGGACCGCCC CGGGGACCCC CGGCTCGTCG GGTACATCGT
41221 GCCCGCGCCC GGCGCCGAAC CGGACGCCGG TGAGCTCCGT GCGGCCCTGG CCCGTACCCT
41281 CCCGCCCCAC ATGGTGCCCT GGGCGCTCGT CCCCCTCCCC GCACTGCCGC TGACGTCCAA
41341 CGGCAAACTG GACAGGGCGG CCCTTCCCGT CCCCGCCGCC CGCGCCGGCG GATCCGGGCA
41401 ACGCCCGGTC ACCCCACAGG AGAAGACACT CTGCGCCCTG TTCGCCGACG TCCTCGGCGT
41461 AACGGAGGTC GCCACGGACG ACGTGTTCTT CGAGCTCGGC GGCCACTCCC TCAACGGCAC
41521 CCGGCTGCTC GCCCGGATCA GGACCGAGTT CGGCACCGAC CTCACCCTCC GCGACCTGTT
41581 CGCCTTCCCC ACCGTCGCCG GCCTTCTCCC GCTCCTGGAC GACAACGGAC GGCAGCACAC
41641 CACCCCGCCG CTGCCTCCGC GCCCGGAGCG CCTCCCCCTG TCGCACGCGC AGCAGCGACT
41701 GTGGTTCCTC GACCAGGTCG AAGGCCCCAG CCCCGCGTAC AACATCCCCA CCGCCGTCCG
41761 GCTCGAAGGC CCGCTCGACA TCCCGGCCCT CGCTGTCGCC CTGCAGGACG TCACCAACCG
41821 CCACGAGCCC TTGCGTACTC TCCTCGCCGA GGACTCCGAA GGCCCCCACC AGGTCATCCT
41881 GCCCCCCGAG GCCGCCCGCC CCGAACTGAC CCACAGCACC GTCGCGCCCG GCGATCTCGC
41941 CGCAGCCCTC GCCGAAGCCG CACGCCGCCC CTTCGACCTC GCCGGTGAGA TCCCACTCAA
42001 AGCCCACCTG TTCGGCTGCG GCCCGGACGA CCACACCCTG CTGCTCCTCG TCCACCACAC
42061 CGCCGGCGAC GGAGCCTCCG TCGAGGTCCT CGTACGCGAT CTCGCCCACG CCTACGGCGC
42121 CCGCCGCGCC GGCGACGCCC CGCACTTCGA GCCGCTGCCC CTGCAGTACG CCGACCACAC
42181 CCTGCGCCGA CGGCACCTGC TGGACGATCC GTCGGACAGC ACACAGCTCG ACCACTGGCG
42241 CGACGCCCTG GCCGGCCTGC CCGAGCAGCT CGAACTGCCC ACCGACCACA CCCGGCCCGC
42301 CGTTCCCACC CGCCGGGGCG AGGCGATCGC CTTCACCGTG CCCGAGCACA CGCACCACAC
42361 GCTGCGGGCC ATGGCCCAGG CCCACGGCGT CACCGTGTTC ATGGTCATGC AGGCCGCGCT
```

```
42421 CGCCGCCCTG CTGTCGCGGC ACGGCGCGGG CCACGACATC CCCCTCGGAA CACCCGTCGC
42481 GGGCCGCTCC GACGACGGCA CGGAAGACCT CGTCGGGTTC TTCGTCAACA CGCTCGTACT
42541 GCGCAACGAC GTCTCCGGGG ACCCGACGTT CGCGGAACTC GTGTCGCGGG TGCGGGCCGC
42601 CAACCTGGAC GCGTACGCCT ACCAGGACGT TCCCTTCGAG CGTCTCGTCG ACGTACTCAA
42661 ACCGGAGCGG TCCCTGTCCT GGCACCCGCT CTTCCAGATC ATGATCGCGT ACAACGGCCC
42721 GGCGACGAAC GACACCGCCG ACGGGTCCCG CTTCGCGGGC CTCACCAGCC GCGTCCATGC
42781 CGTCCACACC GGCATGTCCA AGTTCGACCT GTCGTTCTTC CTCACCGAGC ACGCGGACGG
42841 CCTCGGCATC GACGGCGCTC TCGAGTTCAG CACCGATCTC TTCACGCGGA TCACCGCGGA
42901 GCGCCTGGTC CAGCGCTACC TCACCGTCCT GGAGCAAGCC GCCGGAGCAC CGGACCGCCC
42961 CATCAGTTCG TACGAACTCC TCGGCGACGA CGAACGCGCA CTCCTCGCCC AATGGAACGA
43021 CACCGCCCAC CCCACCCCCC CAGGCACGGT GCTCGATCTC CTCGAAAGCC GTGCGGCGCG
43081 GACCCCCGAC CGGCCGGCCG TCGTCGAGAA CGACCACGTC CTCACCTACG CCGACCTGCA
43141 CACCCGGGCC AACCGGCTCG CCCGCCACCT GATCACCGCC CACGGCGTCG GTCCCGAACG
43201 TCTCGTCGCC GTCGCCCTGC CCCGGTCCGC CGAGCTGCTG GTGGCACTTC TCGCGGTCCT
43261 CAAGACCGGA GCCGCCTACG TCCCTCTCGA CCTCACCCAC CCCGCCGAGC GCACCGCCGT
43321 CGTCCTCGAC GACTGCCGGC CGGCCGTGAT CCTCACCGAC GCCGGTGCGG CCCGTGAACT
43381 GCCGCGGCGC GACATCCCAC AGCTCCGCCT CGACGAACCC GAGGTCCACG CGGCGATCGC
43441 GGAACAACCG GGGGGTCCGG TCACCGACCG GGACCGCACG TGCGTCACTC CGGTCAGCGG
43501 CGAGCACGTG GCATACGTGA TCTACACATC CGGCTCCACG GGCCGGCCCA AGGGTGTGGC
43561 GGTGGAACAC CGTTCACTGG CCGACTTCGT GCGGTACTCC GTGACCGCGT ACCCCGGAGC
43621 CTTCGACGTC ACCCTGCTGC ACAGCCCCGT GACCTTCGAC CTCACCGTGA CCTCGCTGTT
43681 CCCGCCACTG GTCGTCGGTG GCGCCATCCA TGTCGCGGAC CTGACCGAGG CGTGCCCACC
43741 GAGCCTGGCC GCGGCGGGCG GGCCGACGTT CGTCAAGGCC ACACCGAGCC ATCTGCCACT
43801 GCTCACGCAC GAGGCGACAT GGGCCGCGTC CGCGAAGGTG CTGCTCGTCG GGGGCGAGCA
43861 GTTGCTGGGA AGGGAGCTGG ACAAGTGGCG GGCCGGGTCG CCGGAGGCCG TCGTCTTCAA
43921 CGACTACGGC CCCACCGAGG CCACGGTCAA CTGCGTGGAC TTCCGTATCG ATCCGGGACA
43981 ACCGATCGGT GCGGGGCCGG TGGCGATCGG CCGCCCGTTG CGGAACACGC GGGTGTTCGT
44041 GCTCGACGGT GGGTTGCGGG CGGTGCCGGT CGGTGTGGTC GGTGAGCTCC ATGTGGCGGG
44101 CGAGGGGCTG GCGCGGGGTT ATCTCGGGCA GCCGGGTCTG ACGGCGGAGC GGTTCGTGGC
44161 GTGTCCGTTC GGTGATGCCG GGGAGCGGAT GTACCGCACG GGTGACCTGG TGCGGTGGCG
44221 TGCGGATGGG ATGCTGGAGT TCGTCGGCCG GGTCGACGAT CAGGTCAAGG TGCGGGGTTT
44281 CCGGATCGAG CTGGGCGAGG TGGAGGCCGC TGTCGCGGCC TGCCCGGGTG TGGACCGCTC
44341 CGTGGTGGTG GTACGGGAGG ACCGACCGGG AGACCGCCGG CTGGTGGCGT ATGTGACGGC
44401 CGCCGGTGAC GAGGCGGAGG GGCTGGCACC GCTGATCGTG GAGACGGCCG CGGGCCGTCT
44461 GCCCGGGTAC ATGGTGCCGT CGGCCGTGGT CGTACTGGAC GAGATTCCCC TGACGCCGAA
44521 CGGCAAGGTG GACCGTGCCG CGCTGCCCGC GCCGCGCGTC GCCCCGGCCG CGGAGTTCCG
44581 CGTCACCGGA TCACCCCGTG AAGAGGCTCT GTGCGCCCTG TTCGCGGAAG TGCTGGGCGT
44641 GGAACGGGTC GGCGTGGACG ACGGGTTCTT CGACCTCGGC GGAGACAGCA TTCTGTCCAT
44701 TCAACTGGTG GCGCGGGCGC GCCGGGCGGG TCTGGAGGTG TCGGTGCGGG ACGTTTTCGA
44761 GCACCGCACC GTACGGGCGC TGGCCGGTGT GGTGCGGGAG TCCGGAGGCG TCGCTGCCGC
44821 CGTCGTGGAC TCCGGTGTGG GTGCGGTGGA GCGGTGGCCG GTGGTGGAGT GGCTGGCGGA
44881 GCGTGGTGGC GGTGGGCTCG GCGGTGCGGT CAGGGCCTTC AACCAGTCCG TCGTGGTCGC
44941 CACACCGGCC GGTATCACCT GGGACGAACT GCGGACGGTC CTGGACGCGG TACGCGAACG
45001 CCACGACGCC TGGCGGCTAC GGGTAGTGGA TTCCGGTGAC GGCGCCTGGT CCCTGCGCGT
45061 CGACGCGCCC GCCCCCGGCG GTGAGCCCGA CTGGATCACC CGGCACGGCA TGGCCAGCGC
45121 CGACCTGGAG GAGCAGGTGA ACGCCGTGCG GGCCGCCGCC GTGGAGGCCC GGAGCCGGCT
45181 CGATCCACTG ACCGGACGGA TGGTCCGCGC GGTATGGCTG GACCGTGGAC CCGACCGCCG
45241 GGGAGTCCTG GTCCTGGTGG CGCACCACCT GGTCGTCGAC GGCGTCTCCT GGCGCATCGT
45301 CCTCGGCGAC CTCGGCGAAG CCTGGACACA GGCACGCGCT GGCGGGCATG TGCGGTTGGA
45361 CACGGTCGGC ACATCGCTGC GCGGCTGGGC GGCGGCGCTG GCGGAACAGG GCCGCCACGG
45421 CGCCCGCGCC ACCGAAGCAA ACCTGTGGGC ACAGATGGTC CACGGCTCGG ACCCTCTGGT
45481 CGGCCCACGC GCGGTGGACC CTTCGGTGGA CGTCTTCGGC GTGGTGGAGT CGGTGGGTTC
45541 ACGGGCGTCG GTGGGGGTGT CGCGTGCCCT GCTGACGGAG GTCCCGTCGG TCCTGGGTGT
45601 GGGCGTGCAG GAAGTGCTGC TGGCGGCATT CGGCCTGGCA GTGACGCGCT GGCGCGGCCG
45661 CGGCGGAAGC GTCGTCGTGG ACGTCGAGGG TCACGGCCGC AACGAAGACG CCGTACCCGG
45721 CGCGGACCTC TCCCGCACCG TGGGGTGGTT CACCAGCATC TACCCCGTCC GCCTCCCCCT
45781 CGAGCCGGCG GCCTGGGACG AGATACGCGC CGGCGGTCCC GCCGTCGGAC GCACCGTCCG
45841 CGAGATCAAG GAATGCCTCC GCACCCTGCC CGACCAGGGC CTGGGCTACG GCATCCTGCG
45901 CTACCTCGAC CCCGAAAACG GACCCGCCCT CGCCCAGCAC CCCACCCCGC ACTTCGGCTT
45961 CAACTACCTC GGACGGGTCT CGGTCTCGGC GGACGCTGCC TCACTGGACG AAGGCGACGC
46021 CCATGCCGAC GGGCTCGGCG GCCTCGTCGG CGGCAGGGCA GCGGCGGACT CCGACGAGGA
46081 ACAGTGGGCC GACTGGGTTC CGGTGTCGGG TCCGTTCGCG GTGGGCGCGG GTCAGGACCC
46141 CGTTCTGCCG GTGGCCCACG CGGTGGAGTT CAACGCGATC ACCCTGGACA CACCCGACGG
46201 CCCCCGCCTC AGCGTGACAT GGTCGTGGCC GACGACACTG CTGTCCGAAT CCCGGATACG
46261 AGAACTCGCC CGCTTCTGGG ACGAAGCCCT CGAAGGGCTG GTCGCACACG CCCGCCGTCC
```

```
46321 CGACGCGGGC GGACTGACCC CCTCGGACCT GCCGCTGGTC GCCCTCGACC ACGCGGAACT
46381 GGAGGCCCTG CAGGCCGACG TCACCGGTGG CGTGCACGAC ATCCTGCCCG TATCACCGCT
46441 TCAGGAAGGA CTGCTCTTCC ACAGCTCCTT CGCCGCCGAC GGGGTCGACG TCTACGTGGG
46501 ACAACTCACG TTCGACCTGA CCGGACCAGT CGACGCCGAC CACCTGCACG CCGTGGTCGA
46561 AAGCCTGGTG ACACGCCACG ACGTCCTGCG CACCGGCTAC CGCCAGGCAC AGTCCGGCGA
46621 ATGGATCGCC GTCGTGGCAC GACAAGTCCA CACCCCCTGG CAGTACATCC ACACACTCGA
46681 CACGGACGCC GACACCCTCA CAAACGACGA GCGCTGGCGG CCGTTCGACA TGACGCAGGG
46741 CCCACTCGCA CGATTCACCC TCGCACGCAT CAACGACACC CACTTCCGCT TCATCGTCAC
46801 GTACCACCAC GTCATCCTCG ACGGCTGGTC CGTGGCGGTT CTCATACGCG AACTCTTCAC
46861 CACCTATCGC GACACCGCCC TCGGCCGCCG GCCGGAGGTT CCGTACTCCC CACCGCGCCG
46921 TGACTTCATG GCGTGGCTCG CCGAACGCGA CCAGACCGCT GCGGGACAGG CATGGCGTTC
46981 CGCGCTGGCC GGACTCGCGG AGCCCACAGT GCTCGCCCTC GGAACGGAGG GCAGTGGGGT
47041 GATTCCCGAA GTCCTTGAGG AAGAGATCAG CGAGGAACTG ACCTCGGAAC TGGTGGCGTG
47101 GGCGCGTGGG CGTGGTGTGA CGGTCGCGTC GGTGGTGCAG GCGGCCTGGG CGTTGGTGCT
47161 GGGGCGGCTG GTGGGCCGGG ACGACGTGGT GTTCGGCCTG ACCGTGTCGG GCCGGCCCGC
47221 CGAAGTGGCG GGTGTGGAGG ACATGGTCGG TCTGTTCGTG AACACCATTC CGTTGCGGGC
47281 CCGGATGGAC CCGGCGGAGT CACTGGGCGC CTTCGTGGAG CGGCTGCAGC GGGAACAGAC
47341 GGAACTGCTC GAGCACCAGC ACGTCCGGCT GGCCGAGGTC CAGCGCTGGG CCGGACACAA
47401 GGAACTCTTC GACGTCGGAA TGGTCTTCGA GAACTACCCG ATGGATTCCC TGCTGCAGGA
47461 TTCACTGTTC CACGGCAGTG GCCTGCAGAT CGACGGAATA CAGGGTGCCG ATGCGACGCA
47521 TTTCGCTTTG AACCTGGCAG TGGTTCCCCT TCCCGCCATG CGATTCCGGC TCGGCTATCG
47581 GCCGGACGTG TTTGACGCGG GTCGGGTGCG TGAGCTGTGG GGTTGGATCG TCCGGGCCTT
47641 GGAGTGCGTG GTCTGCGAGC GTGATGTGCC GGTGTCCGGT GTCGATGTGC TGGGTGCCGG
47701 TGAGCGGGAG ACGCTGCTGG GCTGGGGTGC GGGCGCGGAA CCCGGCGTGC GTGCGCTGCC
47761 GGGTGCGGGT GCGGGTGCGG GTGCGGGGCT GGTCGGGTTG TTCGAGGAGC GGGTGCGGAC
47821 CGACCCGGAC GCGGTGGCCG TGCGCGGCGC GGGAGTGGAA TGGAGTTACG CGGAGCTGAA
47881 CGCGCGGGCG AATGCGGTGG CCCGGTGGCT GATCGGCCGG GGCGTGGGAC CCGAGCGCGG
47941 TGTCGGGGTG GTGATGGACC GCGGCCCGGA CGTGGTGGCC ATGCTCCTCG CGGTCGCCAA
48001 AAGCGGCGGC TTCTACCTGC CCGTCGACCC GCAATGGCCC ACCGAACGCA TCGACTGGGT
48061 ACTCGCCGAC GCCGGCATCG ACCTGGCCGT CGTGGGCGAG AACCTGGCCG CTGCGGTCGA
48121 GGCCGTCCGC GACTGCGAGG TGGTCGACTA CGCGCAGATC GCCCGCGAAA CACGGCTGAA
48181 CGAGCAGGCG GCCACCGACG CCGGTGATGT GACGGACGGG GAGCGCGTGT CGGCTCTGCT
48241 GTCCGGGCAT CCGCTGTATG TCATCTACAC CTCCGGCTCG ACGGGCCTGC CCAAGGGCGT
48301 GGTGGTCACC CACGCCTCGG TCGGCGCCTA TCTGCGGCGC GGCCGCAACG CCTACCGCGG
48361 CGCCGCCGAC GGCCTGGGCC ACGTGCACTC CTCACTCGCG TTCGACCTGA CCGTGACCGT
48421 TCTGTTCACC CCCCTGGTCT CCGGCGGCTG CGTCACCCTC GGCGATCTCG ACGACACCGC
48481 CAACGGCCTG GGCGCCACCT TCCTCAAGGC CACTCCTTCC CACCTGCCCC TGCTCGGCCA
48541 ACTCGACCGG GTACTCGCCC CCGACGCCAC CCTCCTCCTC GGCGGCGAAG CCCTCACCGC
48601 CGGCGCCCTG CACCACTGGC GCACCCACCA CCCCCACACC ACGGTCATCA ACGCCTACGG
48661 CCCGACCGAA CTCACCGTCA ACTGCGCCGA ATACCGCATC CCCCCCGGCC ACTGCCTCCC
48721 CGACGGCCCC GTCCCCATCG GACGCCCCTT CACCGGCCAC CACCTCTTCG TCCTCGACCC
48781 CGCCCTCCGC CTCACACCCC CCGACACCAT CGGCGAACTG TATGTGGCCG GTGACGGCCT
48841 GGCGCGGGGC TATCTCGGGC GCCCGGACCT GACCGCCGAA CGCTTCGTGG CCTGCCCCTT
48901 CCGCAGCCCC GGCGAACGCA TGTACCGCAC CGGCGACCTC GCACGCTGGC GCAGCGACGG
48961 AACACTCGAA TTCATCGGCC GTGCCGACGA CCAGGTCAAG ATCCGCGGCT TCCGCATCGA
49021 ACTCGGCGAA GTCGAGGCGG CTGTCGCGGC GCATCCGCAC GTGGCGCGGG CCATCGCCGT
49081 CGTACGCGAG GACCGGCCCG GCGACCAGCG CCTGGTCGCG TACGTGACAG GCAGCGACCC
49141 GAGCGGCCTG TCCTCGGCGG TGACGGACAC CGTCGCCGGC CGCCTGCCCG CGTACATGGT
49201 GCCGTCGGCC GTCGTCGTAC TGGACCAGAT CCCCCTCACC CCCAACGGCA AGGTCGACCG
49261 CGCCGCCCTC CCCGCGCCCG GGACCGCCTC CGGAACCACC TCCCGAGCAC CCGGCACAGC
49321 CCGTGAAGAG ATCCTGTGCA CCCTGTTCGC CGACGTACTC GGTCTGGATC AGGTCGGCGT
49381 GGACGAGGAC TTCTTCGACC TCGGCGGCCA TTCCCTGCTC GCCACCCGCC TCACCTCACG
49441 GATCCGGTCG GCCCTCGGCA TCGACCTCGG TGTCCGAGCC CTCTTCAAAG CCCCGACCGT
49501 CGGCCGCCTG GACCAGCTGC TCCAGCAACA GACCACCAGC CTCCGGGCAC CCCTGGTCGC
49561 GCGGGAGCGC ACCGGTTGTG AGCCGCTGTC GTTCGCGCAG CAGCGCCTGT GGTTCCTCCA
49621 CCAGCTCGAA GGCCCCAACG CCGCGTACAA CATCCCCATG GCTCTGCGAC TCACCGGCCG
49681 CCTGGACCTG ACCGCGCTGG AAGCGGCCCT GACGGATGTG ATCGCCCGCC ACGAAAGCCT
49741 GCGAACGGTC ATCGCCCAGG ACGATTCGGG CGGCGTGTGG CAGAACATCC TGCCCACCGA
49801 CGACACCCGC ACCCACCTCA CCCTCGACAC CATGCCGGTC GACGCGCACA CCCTGCAGAA
49861 TCGGGTGGAC GAGGCCGCCC GCCATCCGTT CGACCTCACC ACCGAGATCC CCCTCCGCGC
49921 CACCGTCTTC CGCGTCACCG ACGACGAGCA CGTCCTCCTG CTCGTGCTCC ACCACATCGC
49981 CGGCGACGGC TGGTCCATGG CCCCCCTGGC CCACGACCTG TCCGCCGCCT ACACCGTCAG
50041 ACTCGAGCAC CACGCACCGC AACTGCCCGC TCTGGCCGTC CAATACGCCG ACTACGCCGC
50101 CTGGCAACGC GACGTCCTGG GCACCGAGAA CAACACATCG AGCCAACTCT CCACCCAACT
50161 CGACTACTGG TACAGCAAAC TCGAAGGCCT CCCCGCCGAA CTGACCCTCC CCACCAGTCG
```

```
50221 CGTCCGGCCC GCCGTGGCCT CCCACGCATG CGACCGCGTC GAGTTCACCG TGCCCCACGA
50281 CGTGCACCAA GGCCTGACCG CACTCGCCCG CACCCAGGGC GCCACCGTCT TCATGGTGGT
50341 GCAGGCGGCC CTGGCGGCCC TGCTGTCCCG ACTCGGCGCC GGCACCGACA TCCCCATCGG
50401 CACCCCCATC GCCGGCCGCA CCGACCAGGC GATGGAGAAC CTGATCGGAC TCTTCGTCAA
50461 CACCCTCGTA CTGCGCACCG ACGTCTCCGG GGACCCGACC TTCGCCGAGC TCCTGGCCCG
50521 TGTGCGCACC ACTGCTCTCG ACGCATACGC ACACCAGGAC ATCCCCTTCG AACGCCTGGT
50581 AGAAGCCATC AACCCCGAAC GATCCCTCAC CCGGCACCCC CTCTTCCAGG TCATGCTCGC
50641 CTTCAACAAC ACGGACCGCC GATCCGCGCT CGACGCGCTC GACGCCATGC CCGGCCTTCA
50701 CGCACGACCG GCCGACGTCC TGGCTGTGAC CAGCCCCTAC GATCTCGCGT TCTCGTTCGT
50761 GGAGACACCC GGCAGCACGG AGATGCCCGG CATCCTGGAC TACGCAACCG ACCTGTTCGA
50821 CCGCTCCACG GCCGAGGCCA TGACCGAACG TCTGGTGCGC CTCCTCGCGG AGATCGCCCG
50881 CCGGCCCGAG CTGTCCGTGG GCGACATCGG CATCCTGTCG GCCGACGAGG TGAAGGCCCT
50941 CAGCCCCGAG GCTCCCCCGG CAGCCGAGGA ACTTCACACC TCCACACTGC CTGAGCTGTT
51001 CGAGGAGCAG GTGGCGGCTC GGGGCCATGC GGTCGCGGTG GTGTGCGAAG GAGAGGAGCT
51061 GTCGTACAAG GAGTTGAACG CGCGGGCGAA TCGCCTGGCC AGGGTGCTGA TGGAGCGCGG
51121 CGCAGGCCCC GAACGGTTCG TGGGCGTGGC ACTACCGCGT GGCCTGGACC TCATCGTGGC
51181 ACTCCTGGCC GTGACCAAAA CCGGCGCCGC ATACGTTCCG CTCGACCCCG AATACCCCAC
51241 CGACCGCCTC GCGTACATGG TCACCGACGC CAACCCCACC GCGGTCGTGA CCTCAACGGA
51301 CGTACACATC CCCCTGATCG CCCCCCGCAT CGAGCTCGAC GACGAGGCAA TCCGCACCGA
51361 ACTCGCCGCC GCTCCCGACA CAGCCCCCTG TGTCGGGAGC GGCCCCGCCC ACCCCGCCTA
51421 CGTCATCTAC ACCTCCGGCT CCACCGGTCG CCCCAAGGGC GTCGTCATCA GCCACGCCAA
51481 TGTCGTACGC CTGTTCACCG CATGCTCCGA CAGTTTCGAC TTCGGACCGG ACCACGTCTG
51541 GACGCTCTTC CACTCGTACG CCTTCGACTT CTCGGTCTGG GAGATCTGGG GCGCGCTGCT
51601 TCACGGCGGG CGGCTCGTCG TCGTGCCGTT CGAGGTGACT CGTTCTCCCG CCGAATTCCT
51661 CGCGCTGCTC GCCGAGCAGC AGGTCACGCT GCTGAGCCAG ACACCGTCCG CGTTCCATCA
51721 GCTGACGGAG GCCGCCCGCC AGGAGCCGGC GCGCTGCGCC GGGCTGGCCC TGCGACATGT
51781 GGTCTTCGGC GGCGAGGCGC TCGACCCGTC GCGACTGCGC GACTGGTTCG ACCTGCCGCT
51841 CGGCTCACGG CCGACGCTCG TGAACATGTA CGGCATCACC GAGACCACCG TCCACGTCAC
51901 GGTGCTCCCG CTGGAGGATC GCGCGACGAG TCTTTCCGGC AGCCCGATCG GTCGGCCCTT
51961 GGCCGATCTG CAGGTGTACG TCCTCGACGA ACGGCTCCGC CCGGTGCCCC CAGGCACCGT
52021 CGGCGAGATG TACGTGGCAG GCGCCGGTCT GGCCCGCGGC TATCTGGGAC GCCCCGCTCT
52081 GACCGCCGAG CGGTTCGTGG CCGACCCGAA TTCCCGTTCC GGCGGCCGTC TGTACCGCAC
52141 AGGCGACCTG GCCAAGGTGC GGCCCGACGG GGGACTGGAG TATGTGGGCC GCGGGGACCG
52201 GCAGGTGAAG ATCCGCGGCT TCCGGATCGA ACTCGGCGAG ATCGAGGCCG CGCTGGTCAC
52261 ACACGCGGGT GTCGTCCAGG CGGTGGTCCT GGTGCGGGAC GAGCAGACCG ACGACCAACG
52321 GCTTGTCGCG CACGTGGTGC CCGCGCTGCC GCACCGGGCG CCGACCCTGG CCGAACTCCA
52381 CGAGCACCTC GCGGCGACCC TGCCGGCGTA CATGGTGCCG TCCGCGTACC GGACCCTGGA
52441 CGAGCTGCCG CTGACGGCCA ACGGAAAGCT CGACCGCGCG GCGCTGGCCG GGCAGTGGCA
52501 GGGCGGAACC CGCACCCGGA GACTGCCTCG GACGCCGCAG GAAGAGATCC TGTGCGAGTT
52561 GTTCGCCGAC GTCCTCCGGT TGCCCGCCGC CGGGGCCGAC GACGACTTCT TCGCCCTGGG
52621 AGGCCATTCC CTGCTGGCGA CGCGCCTCCT GTCGGCTGTC AGGGGCACCC TGGGTGTGGA
52681 ACTCGGCATC CGCGACCTCT TCGCCGCGCC CACGCCTGCC GGGCTCGCGA CCGTACTGGC
52741 GGCCTCCGGC ACCGCCCTGC CACCTGTGAC CAGGATCGAC CGGCGCCCTG AACGGCTCCC
52801 GCTGTCCTTC GCACAGCGGC GACTGTGGTT CCTGAGCAAG CTGGAAGGGC CCAGCGCCAC
52861 CTACAACATC CCGGTCGCCG TCCGGCTCAC CGGCGCCCTC GACGTCCCGG CTCTCCGGGC
52921 CGCCCTGGGG GACGTCACCG CACGGCACGA ATCACTGCGT ACGGTCTTCC CCGACGACGG
52981 GGGCGAACCC CGCCAGCTGG TGCTCCCACA CGCCGAACCC CCCTTCCTCA CGCACGAGGT
53041 GACCGTCGGA GAGGTGGCGG AACAGGCGGC GTCCGCCACC GGGTACGCCT TCGACATCAC
53101 CAGCGATACG CCGCTGCGGG CCACCCTGTT GCGCGTCTCA CCGGAGGAAC ACGTCCTCGT
53161 GGTGGTCATC CACCACATCG CCGGCGACGG CTGGTCCATG GGGCCGTTGG TGCGTGACCT
53221 GGTCACCGCC TACCGGGCCC GAACGCGGGG CGACGCCCCG GAGTACACCC CGCTTCCCGT
53281 GCAGTACGCC GACTACGCCC TGTGGCAACA CGCTGTTGCG GGCGACGAGG ACGCCCCGGA
53341 CGGCCGGACG GCGCGTCGGC TCGGGTACTG GCGCGAGATG CTGGCCGGGC TGCCCGAGGA
53401 GCACACGCTG CCCGCCGACC GGCCCCGGCC CGTTCGGTCC TCGCACCGGG GCGGCCGGGT
53461 ACGGTTCGAA CTGCCCGCCG GCGTGCACCG GAGTCTGCTG GCCGTGGCGC GTGACCGTCG
53521 GGCCACGCTG TTCATGGTGG TGCAGGCTGC GCTCGCCGGT CTGTTGTCCC GGCTCGGCGC
53581 GGGCGACGAC ATCCCCATCG GCACCCCGGT CGCCGGGCGG GGCGATGAAG CGCTGGACGA
53641 CGTCGTCGGG TTTTTCGTCA ATACCCTGGT CCTTCGGACG AATCTCGCGG GGGATCCGTC
53701 CTTCGCCGAC CTGGTGGACC GGGTCAGGAC CGCCGACCTC GACGCGTTCG CGCACCAGGA
53761 CGTGCCCTTC GAACGGCTCG TGGAGGCGCT TGCGCCACGG CGTTCCCTCG CCCGCCACCC
53821 GCTGTTCCAG ATCTGGTACA CCCTCACCAA CGCCGACCAG GACATCACCG GCCAGGCACT
53881 CAACGCCCTC CCGGGCCTGA CCGGGGACGA GTACCCGCTG GGGGCCAGTG CCGCCAAGTT
53941 CGACCTGTCG TTCACCTTCA CTGAACACCG CACCCCCGAC GGAGACGCCG CCGGCCTGTC
54001 CGTTCTGCTC GACTACAGCA GCGACCTGTA CGACCACGGC ACTGCCGCCG CACTGGGCCA
54061 CCGGCTGACC GGATTCTTCG CAGCACTGGC CGCCGACCCC ACCGCCCCCC TGGGCACCGT
```

```
54121 CCCGCTCCTC ACCGACGACG AGCGGGACCG CATCCTCGGT GACTGGGGCA GCGGTACGCA
54181 CACCCCGCTG CCCCCGCGCA GCGTGGCCGA GCAGATCGTC CGCCGGGCCG CGCTGGACCC
54241 GGACGCCGTC GCCGTCATCA CCGCGGAAGA GGAACTCTCG TACCGGGAAC TGGAAAGGCT
54301 CAGCGGTGAG ACGGCGCGGC TGCTGGCCGA CCGGGGGATC GGCCGCGAGA GCCTCGTCGC
54361 CGTCGCCCTG CCCCGCACGG CCGGCCTGGT CACCACCCTG CTCGGCGTCC TGCGCACCGG
54421 CGCCGCCTAC CTCCCGCTCG ACACCGGGTA CCCCGCCGAG CGACTCGCGC ACGTGCTCTC
54481 CGACGCCCGT CCCGACCTCG TCCTCACCCA CGCCGGCCTC GCCGGACGGC TGCCGGCCGG
54541 CCTCGCGCCG ACCGTCCTCG TCGACGAGCC GCAGCCGCCC GCCGCAGCCG CCCCCGCGGT
54601 TCCCACGTCC CCGTCGGGCG ACCACCTCGC GTACGTCATC CACACCTCCG GCTCCACCGG
54661 CAGGCCCAAG GGCGTCGCGA TCGCCGAGTC CTCCCTGCGC GCCTTCCTCG CGGACGCGGT
54721 CCGGCGCCAC GACCTGACCC CGCACGACCG GTTGCTCGCG GTGACCACCG TCGGCTTCGA
54781 CATCGCCGGC CTCGAACTGT TCGCCCCGCT CCTCGCCGGT GCCGCGATCG TGCTGGCCGA
54841 CGAGGACGCC GTACGCGACC CCGCCTCGAT CACCTCCCTG TGCGCACGCC ACCACGTCAC
54901 CGTCGTCCAG GCCACGCCCA GTTGGTGGCG GGCCATGCTC GACGGAGCAC CGGCCGACGC
54961 CGCCGCCCGG CTCGAGCACG TACGGATCCT GGTCGGCGGC GAACCGCTGC CCGCCGACCT
55021 GGCCCGTGTC CTGACCGCAA CCGGCGCCGC CGTCACCAAC GTGTACGGAC CCACCGAAGC
55081 CACCATCTGG GCCACCGCCG CCCCACTCAC CGCCGGCGAC GACCGCACAC CCGGCATCGG
55141 CACCCCCCTG GACAACTGGC GCGTCCACAT ACTCGACGCG GCCCTCGGAC CCGTTCCCCC
55201 GGGTGTTCCG GGCGAGATCC ACATCGCCGG GTCCGGGCTC GCCCGCGGCT ATCTGCGCCG
55261 CCCGGACCTC ACCGCCGAAC GCTTCGTCGC CAACCCGTTC GCCCCCGGCG AGCGGATGTA
55321 CCGCACCGGC GACCTCGGCC GGTTCCGCCC GGACGGCACG CTCGAACACC TCGGCCGCGT
55381 GGACGACCAG GTCAAGGTAC GGGGCTTCCG CATCGAACTC GGCGACGTCG AGGCCGCCCT
55441 CGCCCGGCAT CCCGACGTGG GGCGCGCCGC CGCCGCCGTC CGCCCCGACC ACCGCGGCCA
55501 GGGCCGCCTT GTCGCGTACG TCGTCCCCCG TCCCGGCACC CGGGGACCGG ACGCCGGCGA
55561 ACTGCGCGAG ACGGTACGCG AACTTCTGCC TGACTACATG GTCCCCTCCG CCCAGGTGAC
55621 TCTCACCACC CTGCCTCACA CCCCGAACGG CAAACTCGAC CGCGCCGCGC TGCCCGCCCC
55681 CGTGTTCGGC ACCCCTGCCG GACGCGCCCC CGCCACCCGC GAGGAAAAGA TCCTCGCCGG
55741 GCTCTTCGCG GACATCCTGG GCCTGCCCGA CGTGGGAGCC GACAGCGGCT TCTTCGACCT
55801 CGGCGGCGAC AGCGTGCTGT CCATCCAGCT CGTGAGCCGC GCCCGGAGGG AAGGACTGCA
55861 CATCACCGTA CGAGACGTGT TCGAGCACGG GACGGTCGGC GCACTCGCCG CCGCGGCCCT
55921 TCCGGCACCG GCCGACGACG CGGACGACAC CGTCCCCGGC ACGGACGTAC TGCCTTCGAT
55981 CAGCGACGAC GAATTCGAGG AGTTCGAGCT GGAGCTCGGA CTCGAGGGGG AGGAAGAGCA
56041 GTGGTGAACC GCCGGTCGAA GGTAGTCGAG GAGATCCTGC CTGTCTCGGC GCTCCAGGAA
56101 GGACTGCTGT TCCACAGCTC CTTCGCCGCC GCCGACGGAG TCGACGTGTA CGCGGGACAG
56161 CTCGCGTTCG ACCTGGTCGG CGCGGTGGAC ACCGGTCGGC TGCGGGCCGC CGTCGAAAGC
56221 CTCGTGGCGC GGCACGGCGT CCTGCGCTCA AGCTACCGTC AGGCGCGCTC CGGGGAGTGG
56281 GTCGCGGTCG TGGCGCGGCG CGTCGCGACG CCATGGCGCG CCGTCGACGC CCGCGACGGT
56341 GCCACGGACG CTGCCGCCGT GGCCCGGGAG GAACGCTGGC GCCCGTTCGA CCTGGGCCGG
56401 GCCCCGCTGG CTCGGTTCGT GCTCGTACGG ACCGACGACG ACCGTTTCCG GTTCGTGATC
56461 ACGTACCACC ACGTCATCCT CGACGGCTGG TCGCTGCCGG TACTGCTGCG CGAACTCCTT
56521 GCCCTGTACG GAAGCGGCGC CGACCCGTCG GTGCTGCCGC CCGTCCGCCC CTACGGCGAC
56581 TTTCTCCGGT GGGCCGCCGC GCGCGACGAC GCCGCCGCCG AAACCGCCTG GCGCGACGCG
56641 CTCACCGGCC TGGACGAGCC CTCCCTGGTC GCACCCGGCG CTTCCCCCGA CGGCGTCGTG
56701 CCGGCCTCCG TCCACGCCGA ACTCGACAAG GCCGGCACCG AGAACCTCGC CGCCTGGGCC
56761 AGGCACCGCG GCATCACCCA GGCCACCGCC GTCCGCGCCG CGTGGGCCCT CGTTCTCGGC
56821 CAGCACACCG GCCGCGACGA CGTCGTGTTC GGCGTCACCG TCTCCGGACG GCCCGCCGAA
56881 CTCGCCGGCG CCGAGCACAT GGTCGGACTC TTCATCAACA CCGTCCCCCT GCGCACGGTC
56941 CTCGACCCCG CCGACACCCT CGGCACGTTC GCCGCTCGCC TCCAGGCCGA ACAGACCACC
57001 CTCCTCGAAC ACCAGCACGT GCGGCTCTCC GACATCCAGC GCTGGGCCGG ACACAAAGAA
57061 CTCTTCGACA CCATTGTCGT CTTCGAGAAC TACCCCATCG GCCACAGCGG CCCCGGCTCC
57121 ATCCGCACCG ACGACTTCAC CGTCACCGCC ACCGAAGGCT CCGACGCCAC CCACTACCCC
57181 CTCACCCTCA CCGCCGTACC CGGCGAAACC CTGCGCCTCA AGCTCGACCA CCGCCCCGAC
57241 CTCGTCGACA CCACCACCGC CACCGCCCTG CTGCGCCGCG TGACCCGCGT CCTGGAAACC
57301 GCCACCGACG ACACCGGGCA CACCCTCGCC CGCCTCGACC TCCTCGACGA CGACGAACGC
57361 CACCGCCTGC TGCGCGGCTG GAACGACACC ACGCGCGAGC AGCCGCCCAC CTACTACCAC
57421 CAGGAATTCG AGGAACAGGC GCGGAGGCGG CCCCACGACA CGGCCCTTGT CTTCACCAGC
57481 ACCTCCTGGA CGTACGAAGA ACTCAACGAC CGCGCCAACC GGCTCGCCCG CCTGCTCGTC
57541 GCCGCCGGCG CCGGCTCCGA CGACTTCGTC GCGCTCGCCT TCCCCCGTTC CGCGGAATCC
57601 GTCGTCGCCA TCCTCGCCGT ACTCAAAGCG GGCGCCGCCT ACCTGCCGCT CGACATGGAC
57661 CAGCCCGCCG AACGGCTCAC CGGCATCCTC GCCGACGCAC ACCCGACCGT CGTCCTCACG
57721 ACCACCACCG CCACCCCGCT GCCGCACCCC GGCCGCACCC TCGTCCTCGA CAGCCCCACC
57781 ACCGCCCGCG CCCTCGCTGC GGCACCCGCA CACAACCTCA CCGACGCCGA CCGCCGTACC
57841 CCGCTCAACG CCCGCAACGC CGCCTACATC ATCCACACCT CCGGCTCCAC CGGACGCCCC
57901 AAGGGCGTCG TCATCGAACA CCGCAGTCTC GCCAACCTCT TCCACGACCA TCGGCGCGCC
57961 CTCATAGAAC CCCATGCCGC CGGAGGATCA CGGCTCAAGG CCGGCCTCAC CGCCTCCCTC
```

```
58021 TCCTTCGACA CCTCCTGGGA AGGTCTGATC TGCCTGGCCG CCGGCCACGA ACTGCACCTT
58081 ATTGACGACG ACACCCGCCG AGACGCCGAA CGCGTCGCCG AACTCATCGA CCGGCAGCGC
58141 ATCGACGTCA TCGACGTCAC CCCCTCCTTC GCCCAGCAAC TCGTAGAGAC CGGAATCCTC
58201 GACGAGGGCC GCCACCACCC CGCCGCCTTC ATGCTCGGCG GTGAAGGCGT CGACGCGAAA
58261 CTCTGGACCA GGCTCTCCGA CGTCCCCGGC GTCACCTCGT ACAACTACTA CGGCCCCACC
58321 GAATTCACCG TCGACGCCCT CGCCTGCACG GTCGGCATCG CACCCCGCCC CGTCATCGGC
58381 CACCCCCTCG ACAACACGGC CGCCTACATC CTCGACGGCT TCCTGCGTCC CGTACCCGAA
58441 GGCGTCGCCG GCGAGCTCTA CCTCGCCGGC ACCCAGCTCG CCCGCGGCTA CGCCGGCCGG
58501 CCCGGCCTGA CGGCCGAACG CTTCGTGGCC TGCCCCTTCG GCGCGCCGGG CGAACGCATG
58561 TACCGCACCG GCGACCTCGT CCGGCGCAGT CCCGGCGGCG TGGTCGAATA CCTCGGACGC
58621 GTGGACGATC AGATCAAACT CCGCGGCTTC CGCATCGAAC CCGCCGAGAT CGAGCTCGCC
58681 CTGGCCGGCC ACCCCGCCGT CGCCCAGAAC GTCGTCCTCC TGCACCGCTC CGCCACCGGA
58741 GAGGCTCGCC TCGTGGCGTA CGTCGTCCCC GGCACACCCG TCGACCCGCG GGAACTCACC
58801 GGGCACCTCG CCGCCCGGCT GCCCGCGTAC ATGGTGCCCT CGGCTTTCGT TCTCCTCGAC
58861 ACCCTCCCGC TCACCCCCAA CGGCAAACTG GACCGCGGCG CCCTGCCGGA GCCCGCCTTC
58921 GGTACCGCGC CCCGCCCCGA GCGCCCCCGC ACACCCGTCG AGGAGATCCT CTGCGGCCTG
58981 TACGCCGACG TGCTCGGGCT TCCCTCGTTC GGCGCCGACG ACGACTTCTT CGACGCCGGC
59041 GGGCACTCGC TGCTGGCCAG CAAACTCGTC AGCCGTATCC GTACGAACCT GAAAACCGAA
59101 CTCAACGTCC GCGCCCTCTT CGAGCACCGC ACGGTCTCCT CCCTGGCCAC CGCCCTCCAC
59161 CGGGCCGCGC AGGCCGGCCC CGCGCTCACC GCCGGACCGC GCCCCGCACG GATCCCGCTG
59221 TCGTACGCCC AGCGCCGCCT GTGGTTCCTC AACCGGCTCG ACCGCGACAG CGCCGCGTAC
59281 AACATGCCCG TCGCACTCCG CCTGCGTGGC CCCCTGGACA GCACCGCCAT GTGCGCCGCA
59341 CTCACCGACG TCGCCGAACG CCACGAGGCG CTGCGCACCG TGTTCGAGGA GGACCGGGAC
59401 GGTGCCCACC AGATCGTGCT GCCCGCGACC GGCCTCGGCC CTCTGCTCAC CGTGACCGGG
59461 GCCGACGGGA CGACCCTGCG TGCCCTCATC ACCGAGTTCG TACGCAGGCC CTTCGACCTG
59521 GCGGCGGAGA TCCCCTTCCG CGCCGCACTG TTCCGCGTCG GCGACGAGGA ACATGTACTG
59581 GTCGTCGTCC TGCACCACAT CGCCGGGGAC GGCTGGTCCA TGGGACCGCT CGCACGCGAC
59641 GTGGCCGAGG CCTACCGGGC GCGGGCGGCC GGGAGGGCAC CCGACTGGGA ACCGCTGCCC
59701 GTGCAGTACG CCGACTACGC GCTCTGGCAG CGGGAGGTGC TGGGCGCGGA GGACGACGAG
59761 ACCGGCGAAC TCTCCGCCCA ACTCGCCCAC TGGCGCACCC GCCTCGCAGG GGCCCCCGCA
59821 GAACTCACGC TGCCCACCGA CCGCCCACGC CCCGCTGTCG CCTCCACCGC CGGAGACCGC
59881 GTCGAATTCA CCGTGCCCGC CGGACTCCAC CAGGCCCTCG CCGACCTGGC ACGGGCCCAC
59941 GGCGCGACGG TCTTCATGGT CGTCCAGGCC GCCCTCGCCG TCCTGCTGTC ACGTCTCGGC
60001 GCCGGCGACG ACATCCCCAT CGGCACCCCG GTCGCCGGCC GCACCGACGA GGCCACGGAG
60061 GAACTGATCG GGTTCTTCGT CAACACGCTG GTGCTGCGCA CCGACGTGTC CGGCGACCCG
60121 ACGTTCGCCG AACTCCTCGC GCGGGTGCGG GCCACCGACC TCGACGCGTA CGCACACCAG
60181 GACGTGCCAT TCGAACGTCT GGTCGAGGTG TTGAACCCGG AGCGGTCACT GGCACGGCAT
60241 CCACTGTTCC AGGTCATGCT GACGTTCAAC GTCCCGGACA TGGACGGGGT CGGAAGCGCG
60301 CTGGGGAATC TGGGGGAACT GGAGGTCTCC GGTGAGGCCA TCCGGACGGA TCAGACCAAG
60361 GTGGATCTCG CTTTCACGTG CACGGAGATG TACGCCGCGG ACGGTGCGGC CTCGGGAATG
60421 CGCGGGGTGC TGGAATACCG GCTTGATGTG TTCGGTGCGG TACAGGCCCG GGAAACGACG
60481 GAGCGGTTGG TGCGGGTGTT GGAGGGTGTG GTTTCTGGTG GGGGTGGGGT GTCTGTGTCG
60541 GGGGTTGATG TGTTGGGTGT GGGTGAGCGG GAGAGGTTGT TGGGGTGGGG TGTGGGTGGG
60601 CCGGTGCCTG TGGTGCCGGG TGGTGGGTTG GTGGGGTTGT TCGAGGAGCG GGTGCGGGCC
60661 GACGCGGACG CGGTGGCCGT GCGTGGCGCG GGGGTGGTGT GGAGTTATGG GGAGTTGAAT
60721 GCGCGGGTGA ATGTGGTGGC GCGGTGGTTG GTGGGTCGGG GTGTGGGGGC GGAGTGTGGT
60781 GTGGGTGTGG TGATGGGCCG CGGGGTGGAT GTGGTGGTGA TGTTGCTGGC GGTGGCGAAG
60841 GCGGGTGGGT TTTATGTGCC GGTGGATCCG GAGTGGCCGG TGGAGCGGGT GGGGTGGGTG
60901 CTGGCGGATG CCGGGGTGGG GCTGGTTGTG GTGGGGGAGG GGTTGTCGCA TGTGGTGGGG
60961 GATTTTCCTG GGGGTGAGGT TTTCGAGTTT TCGCGGGTTG TTCGTGAGTC GTGTCTTGTG
61021 GAGTTGGTGG CTGCGGATGG GGTTGAGGTT CGGAATGTGA CGGATGGTGA GCGGGCGTCG
61081 CGTCTGTTGC CGGGGCATCC GTTGTATGTG GTTTATACGT CGGGTTCGAC GGGGCGGCCG
61141 AAGGGTGTTG TGGTGACGCA TGCTTCGGTG GGTGGGTATT TGGCGCGTGG TCGGGATGTG
61201 TATGCGGGTG CCGTTGGTGG TGTGGGGTTT GTGCATTCGT CGCTTGCGTT CGATCTGACG
61261 GTGACGGTTC TGTTCACGCC TTTGGTGTCT GGCGGTTGTG TTGTGTTGGG TGAGTTGGAC
61321 GAGTCGGCGC AGGGGGTGGG TGCCTCGTTC GTGAAGGTGA CTCCGTCGCA TCTGGGTTTG
61381 CTGGGTGAGC TGGAGGGTGT GGTGGCGGGG AACGGCATGC TGCTGGTGGG GGGTGAGGCG
61441 TTGTCGGGTG GTGCGCTGCG TGAGTGGCGT GAGCGTAATC CGGGTGTGGT GGTGGTGAAT
61501 GCTTATGGTC CGACGGAGCT GACGGTGAAC TGTGCCGAGT TCCTTATCGC GCCTGGTGAG
61561 GAGGTTCCGG ATGGGCCTGT GCCGATCGGG CGTCCTTTCG CGGGTCAGCG GATGTTTGTT
61621 CTGGATGCGG CGCTGCGGGT GGTGCCGGTC GGTGTGGTGG GTGAGTTGTA TGTGGCGGGT
61681 GTGGGTCTGG CGCGGGGCTA TCTCGGGCGT GCGGGTCTGA CGGCGGAGCG GTTCGTGGCC
61741 TGCCCCTTCG GTGCGCCGGG TGAGCGTATG TACCGTACGG GGGATCTGGT GCGGTGGCGG
61801 GTGGACGGCG CGCTTGAGTT TGTTGGTCGT GCGGATGATC AGGTGAAGGT CCGTGGTTTC
61861 CGTGTGGAGT TGGGTGAGGT GGAGGGTGCT GTTGCGGCGC ATCCTGATGT GGTGCGTGCG
```

```
61921 GTTGTTGTGG TGCGTGAGGA CCGGCCGGGT GATCACCGGT TGGTTGCGTA TGTCACCGGT
61981 GTTGACACGG GTGGACTGTC CTCTGCGGTG ATGCGTGCCG TTGCTGAGCG TCTGCCTGCG
62041 TACATGGTGC CGTCGGCGGT GGTGGTTCTG GATGAGATCC CGTTGACGCC GAATGGGAAG
62101 GTGGACCGGG CGGCGCTTCC GGTGCCGGGG GTGGAGGCGG GCGCGGGCTA CCGGGCGCCT
62161 GTTTCGCCGC GGGAGGAGGT GTTGTGTGGT CTGTTCGCGG AGGTGCTGGG GCTGGAGCGG
62221 GTGGGGGTGG ACGATGATTT CTTCGGGTTG GGTGGTCATT CTCTTCTGGC GACTCGTCTG
62281 ATTTCGCGTG TCCGTGCGGT GTTGGGTGTT GAGGCGGGTG TGCGGGCGTT GTTCGAGGCG
62341 CCGACGGTGA GCCGTTTGGA GCGGTTGCTG CGGGAGCGGT CGGCTTTGGG GGTGCGGGTG
62401 CCTCTGGTGG CACGGGAGCG GACGGGTCGG GAGCCGTTGT CGTTCGCTCA GCAGCGTCTG
62461 TGGTTCCTTG AGGAACTGGA AGGGCCCGGT GCTGCGTACA ACATTCCGAT GGCGCTGCGT
62521 CTGGCCGGTG TTCTGGACGT CGAAGCGCTG CACCAGGCGC TCATTGATGT CATCGCCCGC
62581 CACGAAAGCC TCCGCACCCT CATCGCGCAG GATGCGGGTA CTGCCTGGCA GCACATCCTG
62641 CCCGTTGACG ACCCTCGCAC CCGTCCCGGT CTCCCTCTTG TGGACATCGG TGCCGACGCC
62701 CTTCAGGAGC GGCTCGACGA AGCCGCCGGC CGGCCCTTCG ATCTCGCGGC CGATCTCCCG
62761 GTCCGGGCCA CAGTCTTCCG CCTCACCGAC AACGACCACA TCCTCCTGGT CGTGGCCCAT
62821 CACGTGGCCT TCGACGCGAT GTCCCGTGTG CCGTTCATCC GGAACGTCAA GCGCGCCTTC
62881 GAGGCCCGTA CGAACGGCGC GGCCCCCGAC TGGAGGCCGC TGCCCGTGCA GTACGCGGAT
62941 TATGCGGCCT GGCAGCGCGA CGTACTCGGC ACGGAGGACG ACGAGTCGAG CGAGCTGTCG
63001 GCCCAGCTCG CCTACTGGCG CACCCAACTA GCCTCACTAC CGGCCGAGTT GGCGCTCCCG
63061 ACGGACCGGG CCCGGCCCGC CGTCGCCTCG TACGAAGGCG GCAAGGTCGA GTTCACCGTC
63121 CCCGCCGGGG TGTATGACGG CCTGGTGGCT CTCGCCCGTG CCGAGGGTGT CACGGTCTTC
63181 ATGGTCGTGC AGGCGGCGCT GGCCGCGCTC CTCTCCCGGC TCGGCGCCGG CGACGACATC
63241 CCCATCGGCA CCCCGATCGC CGGCCGCACC GACCAGGCCA CCGAAGATCT CATCGGCTTC
63301 TTCGTGAACA CCCTCGTCCT GCGCACCGAC GTGTCCGGCG ACCCGACGTT CGCCGAACTC
63361 CTCGCGCGCG TCCGGGCCAC CGACCTCGAC GCCTACGCCC ACCAGGACAT CCCCTTCGAA
63421 CGACTGGTCG AAGCGGTCAA CCCCGAGCGC TCCCTCGCCC GCCACCCCCT CTTCCAGGTC
63481 ATGCTGACCT TCGACAACAC GATTGACCGT GAGGTCACGG AGGGCTTCGC GGGCCTCGGG
63541 GTGGAAGGCC TGCCGCTGGG TGCGGGAGCG GTCAAATTCG ATCTGCTCTT CGGTCTCTCC
63601 GAGGTGGGCG GCGAGCTGCG CGGAGCCGTG GAGTACCGCT GCGATCTCTT CGACCACCCG
63661 ACGGTGGCGC AGCTCGCGGA GCGCCTGGTG CGGGTACTGG AGCGCGTGGC TTCCGACGCT
63721 TCGGTACGCA CGGGTGAACT GCCGGTCGTC GGCGAGGCGG AGCGCGCCCG TGTCCTGACG
63781 GAGTGGAATG ACACGGGCGT CCCCGGTGTG CCGGAAACAT TCCTGGAGTT GTTCGAGGCG
63841 CAGGTCGCGG CCCGGGGTGA CGCGCCGGCG GTCGTGTACG AGGGTGAGGT TCTGTCGTAC
63901 CGGGAACTCG ACGCGCGGGC GAACCGCCTG GCCGGGCTGC TGGTGGGGCG CGGTGCGGGC
63961 CCGGAGCATT TCGTGGGGGT GGCGCTGCCG CGTGGGCTGG ATCTGATCGT GGCCCTGCTG
64021 GCCGTGCTCA AGTCCGGTGC CGCGTACGTT CCCCTGGACC CGGAGTACCC GGCCGAGCGG
64081 CTGGTCCACA TGGTCACCGA CGCCGCCCCC GTCGTGGTCG TGACCTCCAC CGACGTACGT
64141 ACTCTGCGGA CCGTTCCCCG GGTCGAGCTG GACGACGAGG CGACCCGCGC CACCCTGGTC
64201 GCAGCCCCCG CCACAGGGCC CGACGTGAAG ATGTCCGCCT CCCACCCCGC GTACGTGATC
64261 TACACCTCCG GGTCCACGGG CCGCCCCAAG GGCGTCGTCA TCAGCCACGG CAGCCTGGCC
64321 AACTTCCTCG CCTGGGCGCG GGAAGACCTG GGTGCCGAGC GGCTCCGGCA CGTCGTGTTG
64381 TCCACGTCCC TCAGCTTCGA CGTCTCCGTG GTCGAACTCT TCGCCCCGCT GTCCTGCGGC
64441 GGCACCGTCG AGATCGTCCG GAATCTGCTG GCCCTCGTCG ACCGCCCCGG CCGATGGTCC
64501 GCGAGCCTGG TCAGCGGCGT GCCGTCGGCC TTCGCGCAGC TGCTGGAAGC CGGCCTCGAC
64561 CGGGCCGACG TGGGCATGAT CGCCCTGGCC GGCGAGGCGC TGTCCGCTCG CGACGTGCGC
64621 CGCGTCCGCG CTGTGCTGCC CGGGGCCCGC GTGGCCAACT TCTACGGCCC GACCGAAGCC
64681 ACCGTCTACG CCACGGCCTG GTACGGCGAC ACCCCCATGG ACGCCGCGGC CCCCATGGGC
64741 CGGCCCCTGC GCAACACGTG TGTGTATGTG CTGGACGACG GGCTGCGCGT GGTGCCGGTC
64801 GGTGTGGTGG GTGAGCTGTA TGTGGCGGGT GTGGGTCTGG CGCGGGGCTA TCTCGGGCGT
64861 GTGGGTCTGA CGGCGGAGCG GTTTGTGGCG TGTCCGTTCG GTGCGCGGGG TGAGCGTATG
64921 TATCGCACGG GGGATTTGGT GCGGTGGCGG GTGGACGGCA CGCTTGAGTT TGTTGGTCGT
64981 GCGGATGATC AGGTGAAGGT CCGTGGTTTC CGTGTGGAGT TGGGTGAGGT GGAGGGTGCT
65041 GTTGCGGCGC ATCCTGATGT GGTGCGTGCG GTTGTTGTGG TGCGTGAGGA CCGGCCGGGT
65101 GATCACCGGT TGGTTGCGTA TGTCACCGGT GTTGACACGG GTGGACTGTC CTCTGCGGTG
65161 ATGCGTGCCG TTGCTGAGCG TCTGCCTGCG TACATGGTGC CGTCGGCGGT GGTGGTTCTG
65221 GATGAGATCC CGTTGACGCC GAACGGGAAG GTGGACCGGG CGGGTCTTCC GGTGCCGGTG
65281 GTGTCGGTGG CGGGGTTCTG TGCGCCGTCG TCGCCGCGGG AGGAGGTGTT GTGTGGTCTG
65341 TTCGCGGAGG TGCTGGGTGT TGAGCGGGTG GGGGTGGACG ATGGGTTCTT CGATCTGGGC
65401 GGGGACAGCA TTCTGTCGAT TCAGTTGGTG GCGCGGGCTC GTCGGGCGGG TCTGGAGTTG
65461 TCGGTTCGGG ATGTTTTCGA GGGCCGTACG GTACGTGCTC TGGCGGCTGT GGTGCGTGGT
65521 TCGGACGCTG GGGCGGTTGG TGTGGTGGGG GGTGCTGAGA TTGTGCTGCC GGGTGTGGGT
65581 GAGGTGGAGC GGTGGCCGGT GGTGGAGTGG CTGGCGGAGC GTGGTGGGGG GTCGCTGGGT
65641 GGTGTGGTTC GGGGTTTCAA TCAGTCTGTT GTGCTTGCTG TGCCTGCTGG GTTGGTGTGG
65701 GAGGAGTTGC GGGTGTTGTT GGGTGCGGTG CGGGATCGGC ATGAGGCGTG GCGGTTGCGG
65761 GTGCTGGATT CCGGGGCGTT GTGTGTTGAT GGTGTTGTTC CGGATGACGG GTCGTGGATT
```

```
65821 GTCCGGTGTG ACCTGAGCGG TATGGGTGTG GATGGTCAGG TGGATGCTGT GCGGGCTGCG
65881 GCTGTGGAGG CGCGTGCGTG GCTGGATCCG TCGGTGGGCC GGGTGGTGCG GGCGGTGTGG
65941 CTGGAGCGTG GTGGTGATCG TTCGGGGGTG TTGGTGCTGG TGGCGCATCA CCTGGTGGTG
66001 GACGGTGTGT CGTGGCGGGT GGTGCTGGGG GATCTGGCGG AGGGGTGGGC GCAGGTGCGT
66061 TCGGGTGGCC GTGTGGAGTT GGGTGTGGTG GGGACGTCGT TGCGGGGTTG GGCGGCGGCG
66121 TTGGCGGAGC AGGGCCGGCG GGGCGAGCGT GCGGGGGAGG TGGAGTTGTG GTCGCGGATG
66181 GTTCGGGGTG CGGATGTTCT GGTGGGGTCG CGTGCTGTGG ATGGTGCGGT GGATGTTTTC
66241 GGCGGGGTGG TGTCGGTTGA TTCGCGGGCG TCGGTGTCGG TGTCGCGTGC GTTGCTGACG
66301 GAGGTGCCGT CGGTTCTGGG TGTTGGTGTG CAGGAGGTGT TGCTGGCGGC ATTCGGGCTG
66361 GCGGTCGCGC GGTGGCGCGG CCGGGGTGGG CCGGTTGTGG TGGATGTTGA GGGGCACGGG
66421 CGTAATGAGG ACGCTGTGCG GGGTGCTGAT CTGTCTCGTA CTGTCGGTTG GTTCACCAGT
66481 GTGTATCCGG TCCGTGTGCC GGTGGAGTCC GCTTCGTGGG ACGAGGTGCG TGCGGGTGGT
66541 CCGGTGGTGG GCCGTGTGGT GCGTGAGGTG AAGGAGACTC TGCGTTCGCT GCCTGACCAG
66601 GGTCTGGGTT ATGGCATCCT GCGCTATCTC GATCCCGAGC ACGGTCCTGC TCTGGCCCGG
66661 CATGCCACCC CGCAGTTCGG TTTCAACTAC CTCGGCCGCT TCACCACCGG AACCGACGAC
66721 ACCGGTGACG AGGGGATGAC GGACTGGGTC CCCGTGTCAG GGCCGTTCGC GGTGGGAGCC
66781 GGCCAGGACC CCGAACTGCC CGTGGCGCAC GCGGTCGAGT TCAACGCGAT CACGCTGGAC
66841 ACCCCGGAGG GCCCGCGCCT GGGCGTGACA TGGTCGTGGC CGACGACGCT GCTGCCGGAG
66901 TCCCGGATAC GGGAGCTGGC CCGCTACTGG GACGAGGCCC TGGAAGGGCT GGTCGAACAC
66961 GCCCGGCACC CCGAAGCCGG CGGCCTCACG CCGTCCGACG TGACGCTGGT GGAAGTGAAC
67021 CAGGTGGAGC TCGACCGTCT GCAGGCGGGG GTCGCCGGTG GTGCGGAGGA GATTCTGCCG
67081 GTGTCGGCCC TGCAAGAGGG GCTGCTGTTC CACAGCGCGT TGGCCTCTGG TGGGGTGGAC
67141 GTGTATGTGG GGCAGCTGGT GTTCGATCTG GTCGGTCCGG TGGACGTCGA CCGGCTGCGC
67201 GCGGCTGTCG AAGGTCTGGT GGCGCGGCAC GGGGTGCTGC GGTCGGGATA CCGCCAACTG
67261 CGGTCGGGCG AATGGGTTGC GGTCGTCGCA CGACAGGTGG ATCTGCCGTG GCAGTCCATC
67321 GACGTGCGCG ACGGCGGTAT CGACGGGTTG GTGGAAGAGG AGCGCTGGCG CCGGTTCGAC
67381 ATGGGCCGGG GTCCACTGGC GCGCTTCGTG CTCATCCGGA CGCACGACGA TCGTTTCCGG
67441 TTCGTCATCA CGTACCACCA CGTCGTCCTC GACGGCTGGT CCGTCCCGGT GCTGCTGCGT
67501 GAGCTGCTGG CCCTGTACGG CAGCTCGGGG GACGTATCGG TTCTGCCGGG GGTCCGCTCG
67561 TACGGCGATT TCCTGCGATG GGTCGCCGCG CGAGACGCCG CAGCCGCCGA AGGCGCATGG
67621 CGGCGGGCGC TGACGGGCCT GGAGGAGCCG TCGCTCGTCG CGCCAGGCGT TTCCCGAGAC
67681 GGGGTCGTCC CGGCGGCGTT CCACGGTGCG GTCGACGGCG ACCTCTCGCA GAAGATCGTG
67741 GCGTGGGCGC GCGGGCGTGG TGTGACGGTT GCGTCGGTGG TACAGGCGGC GTGGGCCTTG
67801 GTGCTGGGGC GGTTGATGGG TCGGGACGAT GTGGTGTTCG GGGTGACGGT GTCGGGTCGG
67861 CCTGCCGAGG TGGTGGGTGT GGAGGACATG GTCGGTCTGT TCGTGAACAC CATTCCGTTG
67921 CGGGCGCGGC TGGATCCGGC GGAGTCGCTG GGTGGTTTCG TGGAGCGGCT GCAGCGGGAG
67981 CAGACGGAGC TGCTGGAGCA TCAGCATGTC CGGCTGGCGG AAGTCCAGCG GTGGGCCGGG
68041 CACAAGGAAC TCTTCGATGT CGGAATGGTC TTCGACAACT ACCCGGTTTC TTCTGAATCC
68101 CCGGAAGCGG AATTCCAGAT CTCACGAACA GGCGGATACA ACGGAACCCA CTACGCACTG
68161 AACCTCGTTG CTTCCATGCA CGGCCTGGAG CTGGAACTGG AAATCGGTTA TCGGCCGGAT
68221 GTGTTTGATG CGGGTCGGGT GCGTGAGGTG TGGGGATGGT TGGTGCGGGT GTTGGAGGGT
68281 GTGGTTTCTG GTGGGGGTGG GGTGTCTGTG TCGGGGGTTG ATGTGTTGGG TGTGGGTGAG
68341 CGGGAGAGGT TGTTGGGGTG AGGGGTGTGG GTGGGCCGGT GCCTGTGGTG CCGGGTGGTG
68401 GGTTGGTGGG GTTGTTCGAG GAGCGGGTGC GGGCCGACGC GGACGCGGTG GCCGTGCGTG
68461 GCGCGGGGGT GGTGTGGAGT TATGGGGAGT TGAATGCGCG GGTGAATGTG GTGGCGCGGT
68521 GGTTGGTGGG TCGGGGTGTG GGGGCGGAGT GTGGTGTGGG TGTGGTGATG GGCCGCGGGG
68581 TGGATGTGGT GGTGATGTTG CTGGCGGTGG CGAAGGCGGG TGGGTTTTAT GTGCCGGTGG
68641 ATCCGGAGTG GCCGGTGGAG CGGGTGGGGT GGGTGCTGGC GGATGCCGGG GTGGGGCTGG
68701 TTGTGGTGGG GGAGGGGTTG TCGCATGTGG TGGGGGATTT TCCTGGGGGT GAGGTTTTCG
68761 AGTTTTCGCG GGTTGTTCGT GAGTCGTGTC TTGTGGAGTT GGTGGCTGCG GATGGGGTTG
68821 AGGTTCGGAA TGTGACGGAT GGTGAGCGGG CGTCGCGTCT GTTGCCGGGG CATCCGTTGT
68881 ATGTGGTTTA TACGTCGGGT TCGACGGGGC GGCCGAAGGG TGTTGTGGTG ACGCATGCTT
68941 CGGTGGGTGG GTATTTGGCG CGTGGTCGGG ATGTGTATGC GGGTGCCGTT GGTGGTGTGG
69001 GGTTTGTGCA TTCGTCGCTT GCGTTCGATC TGACGGTGAC GGTTCTGTTC ACGCCTTTGG
69061 TGTCTGGCGG TTGTGTTGTG TTGGGTGAGT TGGACGAGTC GGCGCAGGGG GTGGGTGCCT
69121 CGTTCGTGAA GGTGACTCCG TCGCATCTGG GTTTGCTGGG TGAGCTGGAG GGTGTGGTGG
69181 CGGGGAACGG CATGCTGCTG GTGGGGGGTG AGGCGTTGTC GGGTGGTGCG CTGCGTGAGT
69241 GGCGTGAGCG TAATCCGGGT GTGGTGGTGG TGAATGCTTA TGGTCCGACG GAGCTGACGG
69301 TGAACTGTGC CGAGTTCCTT ATCGCGCCTG GTGAGGAGGT TCCGGATGGG CCTGTGCCGA
69361 TCGGGCGTCC TTTCGCGGGT CAGCGGATGT TTGTTCTGGA TGCGGCGCTG CGGGTGGTGC
69421 CGGTCGGTGT GGTGGGTGAG TTGTATGTGG CGGGTGTGGG TCTGGCGCGG GGCTATCTCG
69481 GGCGTGTGGG TCTGACGGCG GAGCGGTTTG TGGCGTGTCC GTTCGGTGTG CCGGGTGAGC
69541 GTATGTATCG CACGGGGGAT TTGGTGCGGT GGCGGGTGGA CGGCGCGCTT GAGTTCGTTG
69601 GCCGTGCGGA TGATCAGGTG AAGGTCCGTG GTTTCCGTGT GGAGTTGGGT GAGGTGGAGG
69661 GTGCTGTTGC GGCGCATCCT GATGTGGTGC GTGCGGTTGT TGTGGTGCGT GAGGACCGGC
```

```
69721 CGGGTGATCA CCGGTTGGTG GCTTACGTGA CTGCGGGTGG TGTTGGTGGG GATGGTCTTC
69781 GTTCCGCGAT CTCTGGTTTG GTGGCTGAGC GTCTGCCTGC GTACATGGTG CCGTCGGCGG
69841 TGGTGGTTCT GGATGAGATC CCGTTGACGC CGAACGGGAA GGTGGACCGG GCGGCGCTTC
69901 CGGTGCCGGA GGTGGAGGCG GGCACGGGCT ACCGGGCGCC TGTTTCGCCG CGGGAGGAGG
69961 TGTTGTGTGG TCTGTTCGCG GAGGTGCTGG GTGTTGAGCG GGTGGGGGTG GACGATGACT
70021 TCTTCGAGTT GGGTGGTCAT TCTCTTCTGG CGACTCGTCT GATTTCGCGT GTCCGTGCGG
70081 TGTTGGGTGT TGAGGCGGGT GTGCGGGCGT TGTTCGAGGC GCCGACGGTG AGCCGTCTGG
70141 AGCGGTTGCT CCGGGAGCGG TCGGGTTTGG GGGTGCGGGT GCCTCTGGTG GCACGGGAGC
70201 GGACGGGTCG GGAGCCGTTG TCGTTCGCTC AGCAGCGTCT GTGGTTCCTT GAGGAACTCG
70261 AAGGGCCCGG TGCTGCGTAC AACATTCCGA TGGCGCTGCG TCTGGCCGGT GTTCTGGACG
70321 TCGAAGCGCT GCACCAGGCG CTCATTGATG TCATCGCCCG CCATGAAAGC CTCCGCACCC
70381 TCATCGCGCA GGATGCGGGT ACTGCCTGGC AGCACATCCT GCCCGTTGAC GACCCTCGCA
70441 CCCGTCCCGG TCTCCCTCTT GTGGACATCG GTGCCGACGC CCTTCAGGAG CGGCTCGACG
70501 AAGCCGCCGG CCGGCCCTTC GACCTCGCGG CCGATCTCCC GGTCCGGGCC ACAGTCTTCC
70561 GCCTCACCGA CAACGACCAC ATCCTCCTGC TGGTCCTGCA CCACATCGCC GGCGACGGCT
70621 GGTCGATGGG CCCGCTCGCC CGCGATCTCT CCACGGCGTA CAGCGCACGC GCCGCAGGAG
70681 CCGCCTCGGC CTGGCGGCCC CTCTCCGTGC AGTACGCGGA TTATGCGGCC TGGCAGCGCG
70741 ACGTACTCGG CACGGAGGAC GACGAGTCGA GCGAGCTGTC GGCCCAGCTC GCCTACTGGC
70801 GCACCCAACT AGCGTCACTC CCAGCCGAGT TGGCGCTCCC GACGGACCGG GCCCGGCCCG
70861 CCGTCGCCAC CTACCGGGGC GGACGCATCG AGTTCACCAT CCCCGCCGAC GTCCACCGCA
70921 GCCTCGCCGA CCTCGCCCGT GCCGAGGGTG TCACGGTCTT CATGGTCGTG CAGGCGGCGC
70981 TGGCCGCGCT CCTCTCCCGG CTCGGCGCCG GCGACGACAT CCCCATCGGC ACCCCGATCG
71041 CCGGCCGCAC CGACCAGGCC ACCGAAGATC TCATCGGCTT CTTCGTGAAC ACCCTCGTCC
71101 TGCGCACCGA CGTCTCCGGC GACCCGACGT TCGCCGAACT CCTCGCGCGC GTCCGGGCCA
71161 CCGACCTCGA CGCCTACGCC CACCAGGACA TCCCCTTCGA ACGACTGGTC GAAGCGGTCA
71221 ACCCCGAGCG CTCCCTCGCC CGCCACCCCC TCTTCCAGGT CATGCTCGCC TTCAACAACG
71281 CCGAGACGAG CACCCCGCTG CCCATGGCCG AAGGCCTGGC TGCCTCCCGG CAGGACATCG
71341 AACCGGGCGT GGCGAAATTC GATCTGGCCC TGTATTGCAA CGAATCCCGC GGTGAGACGG
71401 GCGACCACCA GGGCATCAGA AGTGTCTTCG AGTACCGCCG CGACCTGTGG GACGAGGACA
71461 CCGTGCGGCA GCTCGCCGAC CGGTTCCTGC ATGTTCTCGC TGCTTTTGCG GCAGCCCCGG
71521 AGCAACGTGC GAGCAGCGTC GACGTGCTCC GGGCGGGCGA GCGCGACCAA CTGCTGCACG
71581 AGTGGAACGA CACGGCTGCC GCTCTCCCCC CGGCACTGCT GCCCCAGCTG TTCGAGGAGC
71641 AGGTGCGGCG CACCCCGCAC GATGTCGCTC TCGTCTCGGG GAACATCCGG CTCACGTACG
71701 CGGAGCTGGA CGCGCGCGCG AACCGCCTGG CCCACTTGCT GCTCGCCCGG GGCGCGGCCC
71761 CCGAGACGTT CGTCGCGGTG GCCCTGCCCC GGACCGAAGA GCTCCTGGTG GCCCTGCTGG
71821 CCGTACAGAA AACAGGTGCC GGACATCTGC CGCTGGATCC CGGCTTCCCG GCCGAGCGGC
71881 TCAGCTACAT GCTGGATGAC GCCCGCCCTG CGGTGGTCCT CACCACGGAG GACATCAGCG
71941 CCCGCATACC CGGCGGAAGC CATGTGGTAC TCGACTCCGA GCAGGTGACC GGCGAGCTCC
72001 ACGACCACCC GGCCACGTCC CCCGCCGGCC GGGGCAACCC CGCCGGCCCG GCGTACGTGA
72061 TCTACACCTC CGGATCCACC GGCCAGCCCA AGGGCGTCGT CGTACCGTCG GCCGCCCTGG
72121 TGAACTTCCT GGCCGACATG GTGCCCAGGC TCGGGCTCCG CGGTGGCGAC CGCCTGCTGT
72181 CCGTGACCAC CGTGGGCTTC GACATCGCGG CCCTCGAGCT CTTCGTCCCG CTACTGAGCG
72241 GCGCCACCGT CGTCCTCGCG GACGGGGAGA CGGTCCGCGA CCCGGCGCTG GCCCGCCAGA
72301 CGTGCGAGGA CCACGGCGTC ACCATGGTCC AGGCGACACC GAGCTGGTGG CACGGCATGC
72361 TCGCCGACGC GGGCGACAGC CTGCGCGGCG TGCACGCCGT CGTGGGCGGT GAGGCCCTGA
72421 GCCCCGGGTT GCGCGACGCG CTGACACGAG GCGCGCGGTC CGTCACGAAC ATGTACGGCC
72481 CGACGGAGAC GACCATCTGG TCCACCAGCG CCGGGCAGGC CGCCGGGGAC AGCGCTCCCC
72541 CTTCGATCGG CACACCCATC CTCAACACTC GCGTGTATGT GCTCGACGCT GCTTTGTGTG
72601 TCGTGCCACC GGGCGTCGCA GGCGAGCTGT ACATCGCGGG CGACGGCCTC GCGCGGGGCT
72661 ATCTCGGGCG TGCGGGTCTG ACGGCGGAGC GGTTCGTGGC CTGCCCCTTC GGTGCGCCGG
72721 GTGAGCGTAT GTACCGTACG GGGGATCTGG TGCGGTGGCG GGTGGACGGC GCGCTTGAGT
72781 TTGTTGGTCG TGCGGATGAT CAGGTGAAGG TCCGTGGTTT CCGTGTGGAG TTGGGTGAGG
72841 TGGAGGGTGC TGTTGCGGCG CATCCTGATG TGGTGCGTGC GGTTGTTGTG GTGCGTGAGG
72901 ACCGGCCGGG TGATCACCGG TTGGTTGCGT ATGTCACCGG TGTTGACACG GGTGGACTGT
72961 CCTCTGCGGT GATGCGTGCC GTTGCTGAGC GTCTGCCTGC GTACATGGTG CCGTCGGCGG
73021 TGGTGGTTCT GGATGAGATC CCGTTGACGC CGAATGGGAA GGTGGACCGG GCGGCGCTTC
73081 CGGTGCCGGG GGTGGAGGCG GGCGCGGGCT ACCGGGCGCC TGTTTCGCCG CGGGAGGAGG
73141 TGTTGTGTGG TCTGTTCGCG GAGGTGCTGG GTGTTGAGCG GGTGGGGGTG GACGATGATT
73201 TCTTCGGGTT GGGTGGTCAT TCTCTTCTGG CGACTCGTCT GATTTCGCGT GTCCGTGCGG
73261 TGTTGGGTGT TGAGGCGGGT GTGCGGGCGT TGTTCGAGGC GCCGACGGTG AGCCGTTTGG
73321 AGCGGTTGCT GCGGGAGCGG TCGGGTTTGG GGGTGCGGGT GCCTCTGGTG GCACGGGAGC
73381 GGACGGGTCG GGAGCCGTTG TCGTTCGCTC AGCAGCGTCT GTGGTTCCTT GAGGAACTGG
73441 AAGGGCCCGG TGCTGCGTAC AACATTCCGA TGGCGCTGCG TCTGGCCGGT GTTCTGGACG
73501 TCGAAGCGCT GCACCAGGCG CTCATTGATG TCATCGCCCG CCACGAAAGC CTCCGCACCC
73561 TCATCGCCCG CGACAGTGAC GGCACGGCCC GGCAGCAGGT GCTGCCCGTC GGTGACCCCG
```

```
73621 CCGCGCGACC GGCTCTTCCG GTCGTACAGA CCGACGCCGA CACCCTCGTC GCGAAACTGA
73681 ACGAGGCCGT CGGCCGCCCC TTCGACCTCA CGGCCGAGAT GCCCCTGCGT GCCACCGTCT
73741 TCCGGGTGGC CGACGAGGAC CACGCGCTGC TGCTGGTGTT CCACCACATC GCCGGCGACG
73801 GCTGGTCGAC GGGCCTGCTC GCCCGCGACC TGTCCACCGC GTACGCAGCC AGGCTCGAAG
73861 GCCGGGACCC CCAACTGCCA CCCCTCCCCG TGCAGTACGC GGACTACGCG GCCTGGCAGC
73921 GCGACGTACT CGGCACGGAG GACGACGAGT CGAGCGAGCT GTCGGCCCAG CTCGCCTACT
73981 GGCGCACCCA ACTTGCCGAC CTCCCAGCCG AGTTGGCCCT CCCGGCGGAC CGGGTCCGGC
74041 CCGCCAGGGC CTCGTACGAA GGAGGCCGGG TCGGCTTCAC CGTCCCCGCC GGGGTCCTCC
74101 GCGACCTCAC GCGCCTGGCC CGTGTCGAGG GTGTCACGGT CTTCATGGTC GTGCAGGCGG
74161 CGCTGGCCGC GCTCCTCTCC CGGCTCGGCG CCGGCGACGA CATCCCCATC GGCACCCCGA
74221 TCGCCGGCCG CACCGACCAG GCCACCGAAG ATCTCATCGG CTTCTTCGTG AACACCCTCG
74281 TCCTGCGCAC CGACGTCTCC GGCGACCCGA CGTTCGCCGA ACTCCTCGCG CGCGTCCGGG
74341 CCACCGACCT CGACGCCTAC GCCCACCAGG ACATCCCCTT CGAACGACTG GTCGAAGCGG
74401 TCAACCCCGA GCGCTCCCTC GCCCGCCACC CCCTCTTCCA GGTCATGCTC GCCTTCGACA
74461 ACACGGCCGA CGGAGGCCCC GTAGAAGACT TCCCCGGACT GTCCGCAGCC GGGCTGCCGT
74521 TGGGTGCGGG CGCGGCGAAG TTCGATCTGC TCTTCGGTCT CTCCGAGGTG GGCGGCGAGC
74581 TGCGCGGAGC CGTGGAGTAC CGCTGCGATC TCTTCGACCA CCCGACGGCC GCACGGATCG
74641 CGGAGCGCCT GGTGCGGGTG CTGGAGCGGG TCGCCGCCGA CGCGTCGGTA CGCCTGGGCG
74701 AGCTGCCCGT GGTGAGCGAC GCCGAGCGGG CCTGCGTCCT GACGGAGTGG AACGACACCG
74761 CCGTCCCCGG CGTGACGGGA ACGCTGTCGG CGCTGTTCGA GGCACGGGCC GCAGCCCGGG
74821 GCGACGCGCC GGCGGTCGTG TACGAGGGTG AAGAACTGTC GTACCGTGAA CTGAACACAC
74881 GCGCCAACCG CCTCGCCCAT GTCCTGGCCG AGCACGGCGC AGGCCCCGAG CGGTTCGTCG
74941 GTGTGGCCCT GCCCCGCAGT CCGGACCTCG TAGTGGCACT GCTGGCGGTC GTGAAATCGG
75001 GCGCGGCCTA CGTACCGCTC GACCCCGAGT ACCCGGCCGA CCGGCTCGCG TACATGGCCG
75061 GCGACGCTGC CCCCGTGGCG GTCCTGACCC GCGGGGACGT CGAACTCCCC GGGTCCGTCC
75121 CGCGGATCGG GCTGGACGAC ACAGAGATCC GCGCGACACT CGCCACCGCC CCCGGCACGA
75181 ACCCCGGCAC GCCGGTGACC GAGGCCCACC CCGCGTACAT GATCTACACC TCCGGATCCA
75241 CCGGCCGCCC CAAGGGCGTC GTCGTCTCCC ACGGCGCCAT CGTCAACCGG CTCGCCTGGA
75301 TGCAGGCGGA GTACCGTCTC GACGCGACCG ATCGTGTCTT GCAGAAGACT CCGGCCGGTT
75361 TCGACGTGTC GGTCTGGGAG TTCTTCTGGC CGCTGCTCGA GGGCGCGGTC CTCGTGTTCG
75421 CCCGGCCCGG CGGCCACCGG GACGCGGCGT ATCTGGCCGG ACTCATCGAG CGCGAGCGCA
75481 TCACCACGGC ACATTTCGTG CCCTCCATGC TGCGCGTCTT CCTCGAAGAG CCCGGCGCGG
75541 CACTCTGCAC CGGACTGAGG CGGGTGATAT GCAGCGGCGA GGCCCTCGGC ACGGACCTGG
75601 CCGTGGACTT CCGCGCGAAA CTGCCCGTCC CCCTGCACAA TCTGTACGGC CCGACCGAAG
75661 CGGCTGTCGA TGTCACCCAC CACGCGTATG AGCCCGCCAC CGGCACGGCC ACGGTCCCCA
75721 TTGGCCGCCC CATCTGGAAC ATCCGCACCT ACGTCCTCGA CGCCGCCCTG CGTCCTGTGC
75781 CACCGGGCGT GCCCGGCGAG CTGTATCTGG CCGGCGCCGG CCTGGCCCGC GGCTACCACG
75841 GCCGCCCGGC ACTGACGGCG GAGCGGTTTG TGGCGTGTCC GTTCGGTGTG CCGGGTGAGC
75901 GTATGTATCG CACGGGGGAT TTGGTGCGGT GGCGGGTGGA CGGCACGCTT GAGTTTGTTG
75961 GTCGTGCGGA TGATCAGGTG AAGGTCCGTG GTTTCCGTGT GGAGTTGGGT GAGGTGGAGG
76021 GTGCTGTTGC GGCGCATCCT GATGTGGTGC GTGCGGTTGT TGTGGTGCGT GAGGACCGGC
76081 CGGGTGATCA CCGGTTGGTG GCTTACGTGA CTGTGGGTGG TGTTGGTGGG GATGGCCTTC
76141 GTTCCGCGAT CTCTGGTCTG GTGGCTGAGC GTCTGCCTGC GTACATGGTG CCGTCGGCGG
76201 TGGTGGTTCT GGATGAGATC CCGTTGACGC CGAACGGGAA GGTGGACCGG GCGGGTCTTC
76261 CGGTGCCGGT GGTGTCGGTG GCGGGGTTCT GTGCGCCGTC GTCGCCGCGG GAGGAGGTGT
76321 TGTGTGGTCT GTTCGCGGAG GTGCTGGGTG TTGAGCGGGT GGGGGTGGAC GATGGGTTCT
76381 TCGATCTGGG CGGGGACAGC ATTCTGTCGA TTCAGTTGGT GGCGCGGGCT CGTCGGGCGG
76441 GTCTGGAGTT GTCGGTTCGG GATGTTTTCG AGGGCCGTAC GGTACGTGCT CTGGCGGCTG
76501 TGGTGCGTGG TTCGGACGCT GGGGCGGTTG GTGTGGTGGG GGGTGCTGAG ATTGTGCTGC
76561 CGGGTGTGGG TGAGGTGGAG CGGTGGCCGG TGGTGGAGTG GCTGGCGGAG CGTGGTGGGG
76621 GGTCGCTGGG TGGTGTGGTT CGGGGTTTCA ATCAGTCTGT TGTGCTTGCT GTGCCTGCTG
76681 GGTTGGTGTG GGAGGAGTTG CGGGTGTTGT TGGGTGCGGT GCGGGATCGG CATGAGGCGT
76741 GGCGGTTGCG GGTGCTGGAT TCCGGGGCGT TGTGTGTTGA TGGTGTTGTT CCGGATGACG
76801 GGTCGTGGAT TGTCCGGTGT GACCTGAGCG GTATGGGTGT GGATGGTCAG GTGGATGCTG
76861 TGCGGGCTGC GGCTGTGGAG GCGCGTGCGT GGCTGGATCC GTCGGTGGGC CGGGTGGTGC
76921 GGGCGGTGTG GCTGGAGCGT GGTGGTGATC GTTCGGGGGT GTTGGTGCTG GTGGCGCATC
76981 ACCTGGTGGT GGACGGTGTG TCGTGGCGGG TGGTGCTGGG GGATCTGGCG GAGGGGTGGG
77041 CGCAGGTGCG TTCGGGTGGC CGTGTGGAGT TGGGTGTGGT GGGGACGTCG TTGCGGGGTT
77101 GGGCGGCGGC GTTGGCGGAG CAGGGCCGGC GGGGCGAGCG TGCGGGGGAG GTGGAGTTGT
77161 GGTCGCGGAT GGTTCGGGGT GCGGATGTTC TGGTGGGGTC GCGTGCTGTG GATGGTGCGG
77221 TGGATGTTTT CGGCGGGGTG GTGTCGGTTG ATTCGCGGGC GTCGGTGTCG GTGTCGCGTG
77281 CGTTGCTGAC GGAGGTGCCG TCGGTTCTGG GTGTTGGTGT GCAGGAGGTG TTGCTGGCGG
77341 CATTCGGGCT GGCGGTCGCG CGGTGGCGCG GCCGGGGTGG GCCGGTTGTG GTGGATGTTG
77401 AGGGGCACGG GCGTAATGAG GACGCTGTGC GGGGCGCTGA TCTGTCTCGT ACTGTCGGTT
77461 GGTTCACCAG TGTGTATCCG GTCCGTGTGC CGGTGGAGTC CGCTTCGTGG GACGAGGTGC
```

```
77521 GTGCGGGCGG TCCGGTGGTG GGCCGTGTGG TGCGTGAGGT GAAGGAGACT CTGCGTTCGC
77581 TGCCTGACCA GGGTCTGGGT TATGGCATCC TGCGCTATCT CGATCCCGAG CACGGTCCTG
77641 CTCTGGCCCG GCATGCCACC CCGCAGTTCG GTTTCAACTA CCTCGGCCGC TTCACCACCG
77701 GAACCGACGA AACCACCACG GCCGACGCCC TCGACCGGGC CCCCGCGTGG AGCCTTCTCG
77761 CCCGCAGCGC CGCCGGCCAG GACCCCGAAC TGCCCGTGGC GCACGCGGTC GAGTTCAACG
77821 CGATCACGCT GGACACCCCG GAGGGCCCGC GCCTGGGCGT GACATGGTCG TGGCCGACGA
77881 CGCTGCTGCC GGAGTCCCGG ATACGGGAGC TGGCCCGCTA CTGGGACGAA GCCCTGGAAG
77941 GGCTGGTCGA ACACGCCCGG CACCCCGAAG CCGGCGGCCT CACGCCGTCC GACGTGGGCC
78001 TCGCGGAACT CTCCTTTGCT GAGATCGAAC TGCTCGAAGA CGACTGGAGG ACACAGGGAT
78061 GACGCAGCGC GCGATGGAGG ACATACTTCC TCTCACTCCG CTGCAGGAGG GACTGCTGTT
78121 CCACAGTGTT TACGACGAGC AGTCCGTCGA CGTGTACACC GTGCAGGTGG TCGTCGACCT
78181 CGAGGGGCCC GTCGACCCCG AAGCACTGCG CGCCGCCGCG GCCGCCCTGC TGCGTCGGCA
78241 CGCCAACCTG CGGGCGGCCT TCCGGTACGA GCGGCTGCAG CGCCCCGTGC AGATCATCCC
78301 GCGCGAGGTT GCGGTGCCGT GGGAGCACAC CGACGTCGCG AAGCTCGAGG GCGCCGAGCA
78361 GAAGGCCGAG ATCGAACGCC TGCTGCACGA CCAGCGGTGG CGCCGCTTCG ATCTGACGGC
78421 TCCGCCCCTG CTGCGGTTCC TGCTCGTGCG CACAGGCCAC GACCGGCACC GTTTCGCGCT
78481 GACTTTCCAT CACATCCTCA TGGACGGCTG GTCGATGCCC GTCCTGCTGC GGGAACTCAT
78541 CACCCTCTAC CGCACCGGCG ACGAGACCGC CCTGCCCTGG GTCCGGCCGT ACCGGGACTA
78601 CCTGGCCTGG ATCTCCCGCC GCGACCGGGA CGAGGCCGGG CGGGCCTGGT CCAAGGCACT
78661 GGCCGGGGTT GACGAGGCCA CCCTCGTCGC CCCGGGTGCC GACCGGGCCG CCGAGCCGCC
78721 GCTGTGGACC GAGTCCCGGC TCGAACCGGA CCTGGCGGCG ACGCTCGCCG CCCGCGCCCG
78781 CGAGTTCGGC GTCACCCTCA ACACCCTCGT CCAGGCCGCC TGGGCGCTCG TCCTCGGCCG
78841 CCTCACCGGC CGCGACGACG TCGTGTTCGG CGTGACCGTG TCCGGCCGGC CGCCGGAGCT
78901 CGCAGGTGTC GAGGACATGG TGGGCCTCTT CATCAACACC GTGCCGCTGC GTGCCGAGCT
78961 GCTGCCGCAC GAGAGCCTCC GGGACTTCAC CGTCCGCCTC CAGCGCGAAC AGATACAGCT
79021 CCTCGACCAC CAGTACGAAC GACTGGCGGT CATCCAGCGG CTCGCCGGCC GGACAGAACT
79081 CTTCGACACG GTGATGGTCT TCGAGAACTA CCCCGTCGCC GCCGCATCCT CCGCCGGCGC
79141 CGACGGCCCC GCGGCCGAAC CCCGGGTCGC CGACGTCCAC GTACGCGACG CCATGCACTA
79201 CCCCCTCGGT CTGCTGGTCC TGCCCGGCCC GCCGCTGCGC CTGCGCTTTG GCCACCGGCC
79261 GAGCGCCCTG CCCGCCGAAC GCGTCACGAC GATCCGCGAC AGCCTCGTGC GAGCCCTGGA
79321 GCTCATGGCC GACCAGCCGG ACCTCGCCGT CGGCAGGGCC GACATCCTCG GCGAGGAGGA
79381 GAAACAGCAT CTCCTCACCG GCCTCAACGA CACCCACCGC GACGTGCCCC CGCTCACCGT
79441 GCCCGGAATG ATCGAGGCCC AGGCGGCCCG CACCCCCGGC AGGCCGGCGG TCCATGCCCG
79501 CGACGGCGAA CTCTCCTACG CCGAACTCAA CGCGCGCGCC AACCGGCTCG CACGCCACCT
79561 CGCCGCGGCC GGCGTGGGCC CCGAGCAGTA CGTCACCCTG CTGCTCCCGC TCTCCGCCCG
79621 CATGGTCGTG GCCGCTCTCG CCGTGATGAA GACCGGCGCC GCGTACGTTC CCGTGGACCC
79681 GGAGTATCCG GCCGACCGCA TCGCGTACAT GCTTGGCGAC ATCGGCCCCG CGCTCGTCCT
79741 CACCGACTCC CGCTCGGCCG CGGCCATGCC CGCCGGCCCG GCCCGCGTCC TCACCCTCGA
79801 CGACGACGCC CTCGACACGG GCGTTCGCGC CCTGCCCGAA CACGACCTCG GCACCGACGG
79861 TATCGCGCCG CTTCCCGACC AGCCCGCGTA CGTCATCTAC ACCTCGGGCT CCACCGGCCG
79921 CCCCAAGGGC GTCGTGATCC TGCACCGTTC CGTCACCGGC TACCTCCTGC GCACGATCGA
79981 GGAATACCCC GAAGCCGCCG GCAAGGCATT CGTGCACTCG CCCGTGTCCT TCGACCTCAC
80041 CGTCGGAGCG CTGTACGCAC CCCTGGTGAG CGGTGGCTGC CTGCGCCTCG GATCGTTCAC
80101 CGACGACAAG ATCCTCGACC TGGGCGAGGA CAGCCCCACC TTCATGAAGG CCACCCCCAG
80161 CCATCTCGCC GTCCTCGACT CCCTCCCCGA CGAGATCTCC CCCACCGGGG CCATCACCCT
80221 CGGCGGTGAG CAACTCCTGA GCGAGACCCT CGACCCGTGG CGCGCCCGCC ACCCCGGCGT
80281 GACCGTCTTC AACGTGTACG GCCCCACCGA GACCACGATC AACTGCGCCG AACACCGCAT
80341 CGCCCCCGGC ACCACCCTGC CTCCCGGCCC CGTCCCCATC GGCCGGCCCC TGTGGAACAC
80401 CCGCCTGTAC GTCCTCGACG GCGGCCTGCG CGTCGTGCCC ACGGGCGTCG CCGGCGAGCT
80461 GTACGTGGCC GGCGCGGGCC TGGCCCGCGG CTATCTCGGA CGCCCCGGCC TGACGGCCGA
80521 ACGCTTCGTG GCCTGCCCCT TCGGCGCACC GGGCGAACGC ATGTACCGCA CCGGTGACCT
80581 GGTGCGGTGG AGAACCGACG GCACGCTGGA GTTCGTCGGC CGCGTCGACG ACCAGGTCAA
80641 GGTACGCGGC TTCCGCATCG AGCTCGGTGA GGTCGAGGCC ACCGTCGCCG CCACCCCCGG
80701 TGTGGCGCGC GCGATCGTCG CTGTCCGCGA GGACCGCCCC GGCGACCAGC GGCTCGTGGC
80761 GTACGTGACA CCTGCCGACG TCGACCCCAC CGGCGGCCTG CCGTCGGCGG TGACCGCCCA
80821 TGCCGCCGCC CGCCTGCCCG CGTACATGGT GCCGTCCGCC GTCGTGGTAC TGCACGAGGT
80881 ACCCCTCACC CCCAACGGCA AGATCAACAG GGCGGCCCTG CCCGCGCCCG AGGCCGTCTC
80941 CGGCGCCGGC TTCCGTGCCC CCGGCACGGC CCGTGAGGAA GTTCTGTGCG GCCTGTTCGC
81001 CGAAGTCCTC GGCCTCGAAC GGGTCGGCAC GGCCGACGAC TTCTTCGAAC TCGGCGGCCA
81061 CTCGCTGCTC GCCACCCGCC TGGTGTCCCG CGTCCGTTCG GTCCTCGGCG TCGAACTCGG
81121 CGTCCGCGCC CTCTTCGACG CCCCTACCCC CGGCCGCCTC GACCGGCTCC TGGGGGAACG
81181 CTCCGGCGCC CCCGTCCGCG CCCCCCTGAC CGCGCGGGAA CGCACCGGGC GGGACCCCCT
81241 GTCGTACGCC CAGCAGCGCC TGTGGTTCCT CCACGAACTC GAGGGCCACG GCGCCACATA
81301 CAACATCCCT CTCGCGCTGC GCCTCACCGG TCCTCTCGAC GTGACCGCCC TCGAAGCCGC
81361 CCTGACGGAT GTCGTCGCCC GCCACGAGAG CCTGCGCACA CTCATCGCCC GGGACGGCAC
```

```
81421 CGGCACCGCG TGGCAGCACA TCCTGCCCAC CGGCGACCCT CGCGCCCGAA TCACCCTTGA
81481 GGCCGTACCC CTGCACAGGG ACGAACTGGC CGGGCGCCTC GCCGAAGCGG CCCGCCACCC
81541 CTTCGACCTC ACCGCCGAGA TCCCCGTCCG CGCCACCGTC TTCCGCACCG AGCGCGACGA
81601 CCACACCCTG CTCGTCGTCA CCCACCACAT CGCAAGCGAC CGTTGGTCCC GCGAGCCGTT
81661 CCTCCGTGAC CTGTCCGCCG CCTACGCAGC CCGGCGCGCA CACTCCGCGC CGGAACTGCC
81721 CCCGCTGTCC GTGCAGTACG CTGACTACGC CGCCTGGCAG CGCGACGTAC TCGGCACCGA
81781 GGACGACGGG ACGAGCGAGA TGGCCGGCCA GCTCGCCCAC TGGCGGGGCA GACTCGCCGG
81841 CCTCCCGCAG GGCCTGGACC TGCCCACCGA CCGCCCCCGA CGCCCCGACG TCGGCCGCCG
81901 CGGCGGCCGG TGCCGGCTGG AGATCCCCGC CGCGCTGCAC CGCGACATCG TCACCCTCGC
81961 CCGCGTCACC AGTACCACCG TGTTCATGGT GGTCCAGGCG GCCCTCGCCG GTCTGCTGTC
82021 GCGGCTGGGC GCGGGCACCG ACATCCCCAT CGGCACGCCG ATCGCGGGCC GCACCGACGA
82081 GGCCACCGAG CACCTCATCG GGTTCTTCGT GAACACCCTC GTCCTGCGCA CCGACGTCTC
82141 CGGCGATCCG ACGTTCGCCG AACTCCTCGC GCGCGTGCGG GCCACCGACC TCGACGCGTA
82201 CGCACACCAG GACGTGCCCT TCGAACGCCT GGTGGAGGTC CTCAACCCGG AACGCTCACT
82261 GCTGCGCCAC CCCCTCTTCC AGATACTGCT CGCCTTCCAG AACACCGAGG ACCGCAGCAT
82321 CTCCGACCGC CCCGGGACCC TGCTGCCCGA CCTGCAGGTC ACCGAACAGC CCCTCGACGC
82381 CGGGACGGCC AAGTTCGACC TCGCGTTCGC GTTCACCGAG CGGCCCCCGG AGAAGGGCGA
82441 ACCCTCCGGC ATCACCGGAA TCGTCGAATA CCACGCCGAC CTGTACGACG AGGGCACCGT
82501 CCGGCAGATC GCGGACTGCT TCGTGCAGTT CCTCGACGCG GCCGTCCACG CCCCGGGCAC
82561 CCGCGTCGAC GCGGTCGGGC TGCTCCCGGA ACACACCCTC CACAAACTGC TGACCCGCAG
82621 CCGCGGCACT GTCACCGGCC TGCCGCCCGC CACCCTGCCC GAGCTGTTCG AGGCCCGGGT
82681 GGCGGCGCAC CCCGGTCACA TCGCGGTCGA GGTCGCCGGC CGCCGGCCCG CCACTACGAC
82741 GTACGACGCA CTGAACCGGC GGGCCAACCG GCTCGCCCGG CTGCTCACCG ACCGGGGCGT
82801 ACGGCCCGAA CAGCGCGTGG CGATCGCCCT GCCCCGCTCC GCGGACCTGG TGACGGCCTG
82861 GCTCGGGATC CTCAAGGCCG GCGCCGTGTG CGTGCCCGTC GACCCCGCCT ACCCCGACGA
82921 CCGCATCGCC CACATGGCCG CCGACGCGGC CCCGGCGCTC CTCATCGCCT CCGCAGCCAC
82981 CCGCGACCGC ATGCTCCCCA CCGGCATCCC CGTACTGGAC CTCGACGACC CGGCCGTCAC
83041 CGCCGCACTC GCCGCCGCGC CCGACGGCAA TCCGCGCGGC ACGGGACTGC TGCCCGCCCA
83101 TCCCGCCTAC GTCATCTACA CCTCCGGCTC CACCGGCACA CCCAAGGGCG TCGTCGTCAC
83161 CCACGAAGGC ATCCCGGCGC TGGCCGCCAC CCAGCAGGAG GCACTGCGCG CGGGCCCCGG
83221 AGACCGGGTC CTGCAACTGG TGTCGACCAG CTTCGACGCC TCCGTCTGGG ACCTGTGCTC
83281 CGCGCTGCTG TCGGGCGCGA CCCTCGTCCT CGCCCCGGAC GCGGACCTCT TCGGTGACGA
83341 ACTCGCCGCC GCGCTCACCG CACACCGCAT CACGCACGTC ACCCTGCCCC CGGCCGCGCT
83401 GGCCGCTGTC CCGGCAGGCG CGGCACCCCC CCGGCTGACG GTCACCGTCA CCGGCGACGT
83461 GTGCGGACCC CAACTCGTCG ACCGCTGGGC CGGTGGCGAA CGGCGGATCC TCAACGGCTA
83521 CGGGCCCACC GAGGTCACCG TCGGCGCCAC CTACGCCGTG TGCGAACGGA CCGGTGACGG
83581 CGCGCCCGTG CCGATCGGCG CACCCTGGCC CGACCAGCGT GTGTACGTCC TCGAACACCG
83641 GCTCCGGCCC GTACCCGCCG GCTGCGTCGG CGAGATCTAC GTCGCCGGGG CCGGACTGGC
83701 CCGCGGCTAT CTGGGCCGCC CCGGACAGAC CGCCGAACGC TTCGTCGCCG ACCCCTTCGG
83761 CGCCCCCGGC GAGCGCATGT ACCGCACCGG TGACCTGGCC CGCCGCCGCA GCGACGGCCA
83821 CCTGCTGTTC GAGGGACGCG CCGACACGCA GGTCAAAATC CGCGGCTTCC GCGTCGAACT
83881 CGCCGAGATC GAGGCGGCCC TCGCATCGCA CCCCGGCGTC GAGGACGCGG TGGTCACCGT
83941 GTACGACGAC GGGCTCGGCG ACCAGCGGCT CGTCGCGTAC GTCACCGGCG GCCCCGGCAC
84001 ACCGTCGGCC GCCGCGCTGC GCGCCCACCT GGCGTCCCGG CTGCCCCGGC ACATGGTGCC
84061 CGGTGACGTC CTCACCCTGG ACGCCCTGCC GCTCACCGCC AACGGCAAGG TGGACCGCAC
84121 GGCGCTGCCC GGCCCCGGCA CCCAGACCGC CGCCCCCGGG CGCGCACCCC AGTCGCCGCA
84181 GGAACGGGTG CTGTGCGCCT TGTTCGCCGA CGTGCTCGGC CGGGAGACCG TCGGCGTGGA
84241 CGAGGGGTTC TTCGACCTGG GCGGTCACTC GCTGCTCGCC ACTCGCCTCG CGGCCCGGGT
84301 CCGCGCGGCG CTGGGCGTGG AGATCTCCGT GCGCACCCTG TTCGAGGCGC CGACCCCTGC
84361 CCTGCTCGCG TCGGCGTGCA CGGCGGACGC CGCGGCGTAC GACCCGTTCG AGACGGTGCT
84421 GCCGCTGCGG CGCACGGGCA GCCGGCCACC GCTGTTCTGC GTCCACGCCG GAATGGGCCT
84481 GAGCTGGGCG TACGCCGGCC TGCTCAGCCA TCTGGACGCG GACGTGCCGG TTTACGGACT
84541 GCAGGCCCGG AGGCTCACCG CGCCCGGCGG GCTGCCCGGG AGCGTCGAGG AGATGGCTGA
84601 GGACTACGCC GGTGAGATCC GGCGCCTGTG CCCGGATGGG CCGTACCGGC TGCTCGGCTG
84661 GTCCTTCGGC GGCACGGTCG CCCACGCCGT CGCGACCCGC CTGCAACAGC AGGGCCACAC
84721 CGTCGAACTC CTCGCCGTCC TCGACGCCTA CCCCGTCACC GGGGCCCGGC CCGACGCCGA
84781 GGTGGACGAA CAGCGCATCG TCGCCGACTA CCTCGCCCAG CTCGGTTCCC CCGTCGCCCC
84841 CGAGCGCCTC GAGGGCGACG CGTGGCTCCC GGAGTTCCTC GAGTTCGTAC GGCGCACCGA
84901 CGGGCCCGCG AGGGACTTCG ACGCCGGGCG GATCCTCGCG ATGAAGGACG TCTTCCTCAA
84961 CAACGCCCGG CTCACCCGCC GTTTCACACC CGGCGTGTTC ACCGGCGACA TGGTGTTCTT
85021 CGCCTCCGCA CGGCCCGGTT CCGAGCAGGC CGCCGAACGC GTCGGCCTGT GGCACCCCCA
85081 CGTCACCGGC GACCTCGACC TGCACCTGAT CGACTGCGCA CACGAGGAGA TGACCGATCC
85141 AGCCGCACTC ACCCGGATCG GCCCCGTGCT CGCCGCACGG CTGGGCGCCG GCACCTGACC
85201 CCCAGGACCC CACACGGGAC ACCGGACACG GGGGCGCCCC CCTGTCCGTA CACGAAAGGA
85261 AACATACCGC CATGGCCAAC CCCTTCGAGA ACAACGACGG CAGCTACCTC GTACTGGTCA
```

```
85321 ACGACGAGGG CCAGTACTCC CTTTGGCCCG CGTTCGCCGA TGTCCCGGCG GGCTGGACCG
85381 TCACCTTCGG CGAGAGCAGT CGGCAGGAAT GCCTCGACCA CATCAACGAG AACTGGACCG
85441 ATATGCGCCC CAAGAGCCTC ATCCGGCAGA TGGAGAACGA CCGGACGACC GCGGCCTGAC
85501 CCGCAGCCGG ACAGCGGAGA CGGAAGGAGG GCCGACATGA GGGCGACATC CAGGATGATC
85561 CAGGTCAACG GCGCCCGGAT CGCCTGCTCC GACAGCGGCT GCGGTGACCC GGTGCTGATG
85621 ATCGCCGGCA CCGGCAGTAC CGGCCGGGTG TGGGACGCCT ACCAGGTGCC TGACCTGCAC
85681 GCGGCCGGAT TCCGCACCAT CACGTTCACC AATCGCGGCG TACCGCCGTC CGACGAGTGC
85741 GAGCGGGGCT TCACCCTCGC CGACCTCGCC GCCGACACCG CCGCGCTGAT CGAACAGGTG
85801 GCGGGCGGAC CCTGCCGCGT CGTGGGCACG TCCCTGGGCG CCCAGGTGGC CCAGGAAGTC
85861 GCCCTGGCCC GCCCGGACCT GGTGACCCAG GCGGTGTTCA TGGCCACCCG GGGTCGCACC
85921 GACGCGATGC GGGCCGCCGC CACCAGGGCG GCCGCCGCCC TGTACGACAG CGGCGTCGAA
85981 CTGCCCCCCG CCTACGCGGC GGCTGTCCGC GCGCTGCAGA ACCTCTCCCC CCACACCCTC
86041 CGGGACCGCC ATCAGGTCGA GGACTGGCTC CCACTCTTCG AGTACGCCGA ACGGGACGGG
86101 CCGGGGGTCC GTGCGCAGTT GGAACTCGGC CTGCTGCCCG ACCGCCTCGC GGACTACCGG
86161 GACATCACCG TCCCCTGCCT GGTCATCGCG TTCGAGGACG ACGTCGTCAC CCCGCCGTAC
86221 CTGGGCCGCG AAGTGGCCGA CGCGATCCCC GGCGCCCGCT TCGAGACCGT TCCCCGCTGC
86281 GGCCACTACG GCTACCTCGA GGATGCGAGC GCGGTCAACA AGATTCTTCG CGATTTCTTC
86341 CGAACGAGCT GAAAGGCACG ACGACCTTGT CCAGTACCGG CAGAGAGGGG CCCGTCGTGA
86401 CCGGCGAAAC CCGCACCACC ACCTACCTCC CCGGCATGAC CGTGCACGAC TACCACGTGA
86461 CCGTCAAGGA ACAGCACCCG GCGCTCTTCG AGCTCCTGGA CCCCGCACGC CTCGTCGCCG
86521 TCACGGACGA GCCTTGGGTC ACGGAGGGAA ACGAGTTCGA CGACGACCAC GCCGGCCGCG
86581 GCGTCTCCTA CCGCTGTGCC CAGCAGCACG GCGAAGCCCG CCGCACCGGC ATTGAGACGA
86641 TTCTCGGCAT GTTCGCCGGC CCCGGCGGGC TGCGCGACAT GGGCCGTGTC CTCGATGTAC
86701 TCGGAGGCGA AGGCCTGCTC AGCCGCGTGT GGCGGCAACT GGCCGGCGCC GGCGACGGGG
86761 ACTCCGTGCC ACTGGTCACC GGAGACCTCA GCGGCCACAT GGTGGCCGCA GCCCTCCGGT
86821 CCGGCCTGCC CGCCGTACGC CAGCCGGCCG ACCGCATGCT GCAGCGAGAC CACTGCCTGG
86881 ACGGCGTGCT CTTCGCGTAC GGCACTCACC ACGTCGACCG CTCTGTACGC CCCCGCATGC
86941 TGACAGAGGC CTCCCGGGTC CTGGCCCCTG GAGGCCGCGT CGTCCTCCAC GACTTCGCGG
87001 AGGGATCCCC CGAAGAACGC TGGTTCCGCG AAGTCGTCCA CCCCCGCTCC CTCGCGGGCC
87061 ACGCGTACGA CCACTTCACC GCCCACGAGA TGACCGGCTA CCTCGCCGAC GCGGGCTTCA
87121 CCGACATCAC CGTCGGCCCC GTGTACGACC CGATGACCCT GACCGGGGAG ACCGACGAGA
87181 GCGCACTGGC TCGGCTCGTC TCCTACATGA CCTCGATGTA CGGCATCCTG CCCGACGGCG
87241 ACCGGAGCAA CGAGCGGACG GAAGCCGCCC TCCGCGACAT CTTCCGTTTC TCGGCCGGCG
87301 ACCTCCCCGA GGACGTCCCC CGCGACGAGG CGGTCCTGGA ACTTACCGTC CGTCCGCACG
87361 GCAATGCCTT CCGGGCCGAG CTCCCCCGGA TAGCCCTCGT CGCCCACGGA CGCAAACCAT
87421 GACAGCGCAG GACACCCGGA CGACCGGGAG TGACGGTGGC GGCCGGGGCG CCACGTACCA
87481 CGAGAGCCCG ACCTACGGGG AGCTGCTGCG CCTGGAGGAC CTGCTGAACG TCGCGCACCT
87541 GCGCGACGCG GCCGCCCCGG TCCTCTTCCT TGCCACGCAC CAGTCGGCGG AGATCTGGTT
87601 CGGCATCGTG CTGCGCCACC TGGAGGAAAT CCGCGCGGCC CTCACGGACG ACGACCCGGA
87661 CACGGCACTG CATCTGCTGC CGCGACTGCC GGAGATCTTC GAACTGCTCG TCCGCCACTT
87721 CGACATGCTG GCCACGCTGA GTACGGAGGA ATTCGGCAAG ATCCGCGCGG GGCTGGGCAC
87781 GGCGAGCGGC TTCCAGTCGG CGCAGTACCG GGAGATCGAG TTCCTGTGCG GTCTGCGCGA
87841 CCACCGCCAC ATCTCCACAC CGGGCTTCAC GGAAACCGAA CGTCGGCGAC TGCGGGAACG
87901 GGCCCGCCAG CCCTCCGTGG CGGAGGCCTA CGACGCCTTC CGGACCCGAT GCGCCAACGG
87961 GAAGGACGCG GAACGGATCG GGGAAGCGCT CCTGAGGTTC GACGAACGGG TCACCGTCTG
88021 GCGCGCCCGC CACGCGGCCC TGGCGGAACG CTTCCTGGGC CCCCTTGAAG GGACGGCCGG
88081 CACCGCCGGA GCCGACTACT TGTGGCGGGT CACCCGGCAC AGGCTCTTCC CCCCGGAGGC
88141 GTGGGGCGCC GGCTGACGGC ACCGCCCCGG CCCCGGGGAC GGGACAGGCC GGTTCCCGCA
88201 CCCCGGCCCC GGGGGCGGGA AACGGCCTTG CCGTGCCGTC AGAAGGCCGT CAACCGGTCC
88261 CACACGAGGG TCCGAGCCCT TCGTCGAGCA AGCGTCGCCA CTCTGACGTT CGGTCTGTCG
88321 ACGCTCATAC CGGCGGGCAC CGTCACGGCC ACCGGCACCC TGGTCAGGAA GCTGGTGAAG
88381 GAGGCGGGCG AGGAAGCGGC CGGTTGCATC ATGTCCACGC TGACCGAGCC CGCAGTGCAG
88441 GCGATCGAGA ACGTCGCCGC CGACCTGGCG GTTCAGGCCG CAGCCAACGC GGTCGGGCTG
88501 CAGAACGGGA TCGACACCGG GTCAGGCCGT CCACGCCGGC AAGGAGGGGT TCCAGGACGG
88561 AGTCGCGGGT GCGAAGGAAG GACTGCGACT CGCCTCGGTG GACGGCGGTC CGCCGCCGGG
88621 ATCGACGGGC CGGCTGATGG GCGACCTCAA GGCGACCAAG GGCTTTGGCG ACCATCGGGC
88681 GCCAAGACGT GCAAGAACGA CCCCGTGGAC GTCGCCACCG GTGAGATGCT GCTCCCGCAG
88741 ACCGTCCTGG GGCTCCCCGG CGTCCTGCAG CTGGTCCTGG GGCGGACTCA TCCGTGCTCG
88801 GCTGCCGACC GTCGGCAACC TCGACGCCGC CGTGAACTCA TCAGGTTCGC CAGTGCGGTT
88861 CACCTGCGAC GCCGATGGAC GCGTCACCTC CTGGACCGAC TGCAACGACG CCACCTTCCG
88921 TACGTCTACG ACCAGGCCGG CCGGGTGGTG CGGACCGAAG GCCCCGACGG CATCCTCTCC
88981 TCGTCCTGTG CCTATGGAGA GCCGGAACCC GACACCGGGC CGCGCACGAC GCGCAAGGGC
89041 GGCCGGTGAT CCGTCGGAGC TCAACGCTCC GTGGCCGCAC AGCGGTCTGG CACTTCACCT
89101 GGGACGCTCA GGACCGGCTC GCCGAGGCCG CCGACCACTG CTGGGACTGC GCGCGGTTCC
89161 CGCGCCTGAG AGGGGGCGCG CTGACCTTGG TCAGAAGCCT TGCGCGATCA CGAGCGTCAC
```

```
89221 GTTGGCGGGC GACTTGCTCC GCCGAGTTGA AGCCGTACGA AGTGCCGACT GGATCGATGC
89281 GGCCAGCCAT CCTCAGGGTT GCCCTGACAG AGTTTGGTGG ACACGAACCG GAACAACCCG
89341 GACCGCCAGA TGTGGTTGAG TTTTCCGGGA CAGTTGATCC ACACACCGTC CCGCCGCTCA
89401 GTCCCGTTCG GACCATGAGC CTCACCGAAG GACCCCACCT GCACATGAGC ACCCCCGACA
89461 CTCGTCCCCC GCGGTGGCCT TTACCTCCCG CCCCGCCCGC CCATGACCCC GTCCTCTTCG
89521 CGCGGGCGAT GCGGGACATG CGCCTGACGT GGCGTGCCCG CGGGATCCTG GCCGAACTCT
89581 CCGTCGGCTA CGGCCCCGGG CAGGACCCCA CGATCAGCGA GCTGGTCGCG CTCAACCGCG
89641 ACGAGCGTCT GGCTGCAGAG GGCCGCGAGG CCTTCCGCAC GGCGGTCCGT GAGCTGCGCG
89701 GCCTCGGCTA CCTCACTCCG GACGCCACCA CGGCCTCCGG CGTCGGGGAG CGTCTGATCG
89761 TCGATCTCGC CCCGGCCGCG GAAGCCTGGC TGATCCCGCA GCAACCCGGT TTCGGGTTCT
89821 ACGTGGACGG GAGCTGACCG TCCGACTCTC TGCCGGGCGT GGCCGAACGC CGTCCTCGTG
89881 TGACGGGGAC GGCCCGCCTA TCCTGCGGGC ATGGCCCAAC CCATTGAACT CGTCATATTC
89941 GACTGCGACG GCGTACTCGT CGACAGCGAA CGCATCGCGG TGCGCGTGGA CGCACTCGTC
90001 CTGGCCGAGC TGGGGTGGAA TCTCACCGAA GCCGAGATCG TCGACCGGTT CATGGGCCTG
90061 TCGAGCCGGT CGATGACGCG GCAGATCGAG GACCACCTCG GGCGCCGTCT GCCGGCCGAC
90121 TGGGAGGAAG AGTTCAAGCC CCTCTACGAC GAGGCGCTCG CCGCCGAACT CACGCCGGTC
90181 GAGGGCATCG TCGACGCCCT CGACGCGCTC ACGCATCTCC CCACCTGTGT GGCATCCAGC
90241 GGGAGCCACG ACAAGATGCG TTTCACGCTG GGGATGACCG GTCTCCGCCC GCGCTTCGAA
90301 GGCCGCATTT TCAGTGCCAC CGAGGTCGAG CACGGCAAGC CGGCCCCGGA TCTGTTCCTA
90361 CTCGCCGCGC GGAAGATGGG GGTCGTGCCC GAGGCGTGCG CCGTGGTCGA GGACAGTCAG
90421 TACGGTCTTC AGGCAGCCCG GGCCGCGGGC ATGCGAGCCT TCGCCTACGC CGGGGGACTG
90481 ACTCCCGCGG ACCGTCTCGA AGGCCCCGGC ACCGTCGTCT TCGACGACAT GCGCAGACTG
90541 CCCGGCCTCC TCGCGGATCA CTGACCGCCG CCTGGATCAC TCCACTCCAT CGGCCACTGT
```

SEQ ID NO: 2
Nucleotides 36018-36407 of SEQ ID NO: 1

SEQ ID NO: 3
Nucleotides 78059-85198 of SEQ ID NO: 1

SEQ ID NO: 4
Nucleotides 85500-86352 of SEQ ID NO: 1

SEQ ID NO: 5
Nucleotides 85537-86352 of SEQ ID NO: 1

SEQ ID NO: 6
Nucleotides 85537-86352 of SEQ ID NO: 1

SEQ ID NO: 7

MTQRAMEDILPLTPLQEGLLFHSVYDEQSVDVYTVQVVVDLEGPVDPEALRAAAAALLRRHANLRAAFRYERLQRP

VQIIPREVAVPWEHTDVAKLEGAEQKAEIERLLHDQRWRRFDLTAPPLLRFLLVRTGHDRHRFALTFHHILMDGWS

MPVLLRELITLYRTGDETALPWVRPYRDYLAWISRRDRDEAGRAWSKALAGVDEATLVAPGADRAAEPPLWTESRL

EPDLAATLAARAREFGVTLNTLVQAAWALVLGRLTGRDDVVFGVTVSGRPPELAGVEDMVGLFINTVPLRAELLPH

ESLRDFTVRLQREQIQLLDHQYERLAVIQRLAGRTELFDTVMVFENYPVAAASSAGADGPAAEPRVADVHVRDAMH

YPLGLLVLPGPPLRLRFGHRPSALPAERVTTIRDSLVRALELMADQPDLAVGRADILGEEEKQHLLTGLNDTHRDV

PPLTVPGMIEAQAARTPGRPAVHARDGELSYAELNARANRLARHLAAAGVGPEQYVTLLLPLSARMVVAALAVMKT

GAAYVPVDPEYPADRIAYMLGDIGPALVLTDSRSAAAMPAGPARVLTLDDDALDTGVRALPEHDLGTDGIAPLPDQ

PAYVIYTSGSTGRPKGVVILHRSVTGYLLRTIEEYPEAAGKAFVHSPVSFDLTVGALYAPLVSGGCLRLGSFTDDK

ILDLGEDSPTFMKATPSHLAVLDSLPDEISPTGAITLGGEQLLSETLDPWRARHPGVTVFNVYGPTETTINCAEHR

IAPGTTLPPGPVPIGRPLWNTRLYVLDGGLRVVPTGVAGELYVAGAGLARGYLGRPGLTAERFVACPFGAPGERMY

RTGDLVRWRTDGTLEFVGRVDDQVKVRGFRIELGEVEATVAATPGVARAIVAVREDRPGDQRLVAYVTPADVDPTG

GLPSAVTAHAAARLPAYMVPSAVVVLHEVPLTPNGKINRAALPAPEAVSGAGFRAPGTAREEVLCGLFAEVLGLER

VGTADDFFELGGHSLLATRLVSRVRSVLGVELGVRALFDAPTPGRLDRLLGERSGAPVRAPLTARERTGRDPLSYA

QQRLWFLHELEGHGATYNIPLALRLTGPLDVTALEAALTDVVARHESLRTLIARDGTGTAWQHILPTGDPRARITL

EAVPLHRDELAGRLAEAARHPFDLTAEIPVRATVFRTERDDHTLLVVTHHIASDRWSREPFLRDLSAAYAARRAHS

APELPPLSVQYADYAAWQRDVLGTEDDGTSEMAGQLAHWRGRLAGLPQGLDLPTDRPRRPDVGRRGGRCRLEIPAA

LHRDIVTLARVTSTTVFMVVQAALAGLLSRLGAGTDIPIGTPIAGRTDEATEHLIGFFVNTLVLRTDVSGDPTFAE

LLARVRATDLDAYAHQDVPFERLVEVLNPERSLLRHPLFQILLAFQNTEDRSISDRPGTLLPDLQVTEQPLDAGTA

KFDLAFAFTERPPEKGEPSGITGIVEYHADLYDEGTVRQIADCFVQFLDAAVHAPGTRVDAVGLLPEHTLHKLLTR

SRGTVTGLPPATLPELFEARVAAHPGHIAVEVAGRRPATTTYDALNRRANRLARLLTDRGVRPEQRVAIALPRSAD

LVTAWLGILKAGAVCVPVDPAYPDDRIAHMAADAAPALLIASAATRDRMLPTGIPVLDLDDPAVTAALAAAPDGNP

RGTGLLPAHPAYVIYTSGSTGTPKGVVVTHEGIPALAATQQEALRAGPGDRVLQLVSTSFDASVWDLCSALLSGAT

LVLAPDADLFGDELAAALTAHRITHVTLPPAALAAVPAGAAPPRLTVTVTGDVCGPQLVDRWAGGERRILNGYGPT

EVTVGATYAVCERTGDGAPVPIGAPWPDQRVYVLEHRLRPVPAGCVGEIYVAGAGLARGYLGRPGQTAERFVADPF

GAPGERMYRTGDLARRRSDGHLLFEGRADTQVKIRGFRVELAEIEAALASHPGVEDAVVTVYDDGLGDQRLVAYVT

GGPGTPSAAALRAHLASRLPRHMVPGDVLTLDALPLTANGKVDRTALPGPGTQTAAPGRAPQSPQERVLCALFADV

LGRETVGVDEGFFDLGGHSLLATRLAARVRAALGVEISVRTLFEAPTPALLASACTADAAAYDPFETVLPLRRTGS

RPPLFCVHAGMGLSWAYAGLLSHLDADVPVYGLQARRLTAPGGLPGSVEEMAEDYAGEIRRLCPDGPYRLLGWSFG

GTVAHAVATRLQQQGHTVELLAVLDAYPVTGARPDAEVDEQRIVADYLAQLGSPVAPERLEGDAWLPEFLEFVRRT

DGPARDFDAGRILAMKDVFLNNARLTRRFTPGVFTGDMVFFASARPGSEQAAERVGLWHPHVTGDLDLHLIDCAHE

EMTDPAALTRIGPVLAARLGAGT*

SEQ ID NO: 8

See Figure 4, DptH sequence.

SEQ ID NO: 9

MDMQSQRLGVTAAQQSVWLAGQLADDHRLYHCAAYLSLTGSIDPRTLGTAVRRTLDETEALRTRFVPQDGELLQIL

EPGAGQLLLEADFSGDPDPERAAHDWMHAALAAPVRLDRAGTATHALLTLGPSRHLLYFGYHHIALDGYGALLHLR

RLAHVYTALSNGDDPGPCPFGPLAGVLTEEAAYRDSDNHRRDGEFWTRSLAGADEAPGLSEREAGALAVPLRRTVE

LSGERTEKLAASAAATGARWSSLLVAATAAFVRRHAAADDTVIGLPVTARLTGPALRTPCMLANDVPLRLDARLDA

PFAALLADTTRAVGTLARHQRFRGEELHRNLGGVGRTAGLARVTVNVLAYVDNIRFGDCRAVVHELSSGPVRDFHI

NSYGTPGTPDGVQLVFSGNPALYTATDLADHQERFLRFLDAVTADPDLPTGRHRLLSPGTRARLLDDSRGTERPVP

RATLPELFAEQARRTPDAPAVQHDGTVLTYRDLHRSVERAAGRLAGLGLRTEDVVALALPKSAESVAILLGIQRAG

AAYVPLDPTHPAERLARVLDDTRPRYLVTTGHIDGLSHPTPQLAAADLLREGGPEPAPGRPAPGNAAYIIQTSGST

GRPKGVVVTHEGLATLAADQIRRYRTGPDARVLQFISPGFDVFVSELSMTLLSGGCLVIPPDGLTGRHLADFLAAE

AVTTTSLTPGALATMPATDLPHLRTLIVGGEVCPPEIFDQWGRGRDIVNAYGPTETTVEATAWHRDGATHGPVPLG

RPTLNRRGYVLDPALEPVPDGTTGELYLAGEGLARGYVAAPGPTAERFVADPFGPPGSRMYRTGDLVRRRSGGMLE

FVGRADGQVKLRGFRIELGEVQAALTALPGVRQAGVLIREDRPGDPRLVGYIVPAPGAEPDAGELRAALARTLPPH

MVPWALVPLPALPLTSNGKLDRAALPVPAARAGGSGQRPVTPQEKTLCALFADVLGVTEVATDDVFFELGGHSLNG

TRLLARIRTEFGTDLTLRDLFAFPTVAGLLPLLDDNGRQHTTPPLPPRPERLPLSHAQQRLWFLDQVEGPSPAYNI

PTAVRLEGPLDIPALAVALQDVTNRHEPLRTLLAEDSEGPHQVILPPEAARPELTHSTVAPGDLAAALAEAARRPF

DLAGEIPLKAHLFGCGPDDHTLLLLVHHTAGDGASVEVLVRDLAHAYGARRAGDAPHFEPLPLQYADHTLRRRHLL

DDPSDSTQLDHWRDALAGLPEQLELPTDHTRPAVPTRRGEAIAFTVPEHTHHTLRAMAQAHGVTVFMVMQAALAAL

LSRHGAGHDIPLGTPVAGRSDDGTEDLVGFFVNTLVLRNDVSGDPTFAELVSRVRAANLDAYAYQDVPFERLVDVL

KPERSLSWHPLFQIMIAYNGPATNDTADGSRFAGLTSRVHAVHTGMSKFDLSFFLTEHADGLGIDGALEFSTDLFT

RITAERLVQRYLTVLEQAAGAPDRPISSYELLGDDERALLAQWNDTAHPTPPGTVLDLLESRAARTPDRPAVVEND

HVLTYADLHTRANRLARHLITAHGVGPERLVAVALPRSAELLVALLAVLKTGAAYVPLDLTHPAERTAVVLDDCRP

AVILTDAGAARELPRRDIPQLRLDEPEVHAAIAEQPGGPVTDRDRTCVTPVSGEHVAYVIYTSGSTGRPKGVAVEH

RSLADFVRYSVTAYPGAFDVTLLHSPVTFDLTVTSLFPPLVVGGAIHVADLTEACPPSLAAAGGPTFVKATPSHLP

LLTHEATWAASAKVLLVGGEQLLGRELDKWRAGSPEAVVFNDYGPTEATVNCVDFRIDPGQPIGAGPVAIGRPLRN

TRVFVLDGGLRAVPVGVVGELHVAGEGLARGYLGQPGLTAERFVACPFGDAGERMYRTGDLVRWRADGMLEFVGRV

DDQVKVRGFRIELGEVEAAVAACPGVDRSVVVVREDRPGDRRLVAYVTAAGDEAEGLAPLIVETAAGRLPGYMVPS

AVVVLDEIPLTPNGKVDRAALPAPRVAPAAEFRVTGSPREEALCALFAEVLGVERVGVDDGFFDLGGDSILSIQLV

ARARRAGLEVSVRDVFEHRTVRALAGVVRESGGVAAAVVDSGVGAVERWPVVEWLAERGGGGLGGAVRAFNQSVVV

ATPAGITWDELRTVLDAVRERHDAWRLRVVDSGDGAWSLRVDAPAPGGEPDWITRHGMASADLEEQVNAVRAAAVE

ARSRLDPLTGRMVRAVWLDRGPDRRGVLVLVAHHLVVDGVSWRIVLGDLGEAWTQARAGGHVRLDTVGTSLRGWAA

ALAEQGRHGARATEANLWAQMVHGSDPLVGPRAVDPSVDVFGVVESVGSRASVGVSRALLTEVPSVLGVGVQEVLL

AAFGLAVTRWRGRGGSVVVDVEGHGRNEDAVPGADLSRTVGWFTSIYPVRLPLEPAAWDEIRAGGPAVGRTVREIK

ECLRTLPDQGLGYGILRYLDPENGPALAQHPTPHFGFNYLGRVSVSADAASLDEGDAHADGLGGLVGGRAAADSDE

EQWADWVPVSGPFAVGAGQDPVLPVAHAVEFNAITLDTPDGPRLSVTWSWPTTLLSESRIRELARFWDEALEGLVA

HARRPDAGGLTPSDLPLVALDHAELEALQADVTGGVHDILPVSPLQEGLLFHSSFAADGVDVYVGQLTFDLTGPVD

ADHLHAVVESLVTRHDVLRTGYRQAQSGEWIAVVARQVHTPWQYIHTLDTDADTLTNDERWRPFDMTQGPLARFTL

ARINDTHFRFIVTYHHVILDGWSVAVLIRELFTTYRDTALGRRPEVPYSPPRRDFMAWLAERDQTAAGQAWRSALA

GLAEPTVLALGTEGSGVIPEVLEEEISEELTSELVAWARGRGVTVASVVQAAWALVLGRLVGRDDVVFGLTVSGRP

AEVAGVEDMVGLFVNTIPLRARMDPAESLGAFVERLQREQTELLEHQHVRLAEVQRWAGHKELFDVGMVFENYPMD

SLLQDSLFHGSGLQIDGIQGADATHFALNLAVVPLPAMRFRLGYRPDVFDAGRVRELWGWIVRALECVVCERDVPV

SGVDVLGAGERETLLGWGAGAEPGVRALPGAGAGAGAGLVGLFEERVRTDPDAVAVRGAGVEWSYAELNARANAVA

RWLIGRGVGPERGVGVVMDRGPDVVAMLLAVAKSGGFYLPVDPQWPTERIDWVLADAGIDLAVVGENLAAAVEAVR

DCEVVDYAQIARETRLNEQAATDAGDVTDGERVSALLSGHPLYVIYTSGSTGLPKGVVVTHASVGAYLRRGRNAYR

GAADGLGHVHSSLAFDLTVTVLFTPLVSGGCVTLGDLDDTANGLGATFLKATPSHLPLLGQLDRVLAPDATLLLGG

EALTAGALHHWRTHHPHTTVINAYGPTELTVNCAEYRIPPGHCLPDGPVPIGRPFTGHHLFVLDPALRLTPPDTIG

ELYVAGDGLARGYLGRPDLTAERFVACPFRSPGERMYRTGDLARWRSDGTLEFIGRADDQVKIRGFRIELGEVEAA

VAAHPHVARAIAVVREDRPGDQRLVAYVTGSDPSGLSSAVTDTVAGRLPAYMVPSAVVVLDQIPLTPNGKVDRAAL

PAPGTASGTTSRAPGTAREEILCTLFADVLGLDQVGVDEDFFDLGGHSLLATRLTSRIRSALGIDLGVRALFKAPT

VGRLDQLLQQQTTSLRAPLVARERTGCEPLSFAQQRLWFLHQLEGPNAAYNIPMALRLTGRLDLTALEAALTDVIA

RHESLRTVIAQDDSGGVWQNILPTDDTRTHLTLDTMPVDAHTLQNRVDEAARHPFDLTTEIPLRATVFRVTDDEHV

LLLVLHHIAGDGWSMAPLAHDLSAAYTVRLEHHAPQLPALAVQYADYAAWQRDVLGTENNTSSQLSTQLDYWYSKL

EGLPAELTLPTSRVRPAVASHACDRVEFTVPHDVHQGLTALARTQGATVFMVVQAALAALLSRLGAGTDIPIGTPI

AGRTDQAMENLIGLFVNTLVLRTDVSGDPTFAELLARVRTTALDAYAHQDIPFERLVEAINPERSLTRHPLFQVML

AFNNTDRRSALDALDAMPGLHARPADVLAVTSPYDLAFSFVETPGSTEMPGILDYATDLFDRSTAEAMTERLVRLL

AEIARRPELSVGDIGILSADEVKALSPEAPPAAEELHTSTLPELFEEQVAARGHAVAVVCEGEELSYKELNARANR

LARVLMERGAGPERFVGVALPRGLDLIVALLAVTKTGAAYVPLDPEYPTDRLAYMVTDANPTAVVTSTDVHIPLIA

PRIELDDEAIRTELAAAPDTAPCVGSGPAHPAYVIYTSGSTGRPKGVVISHANVVRLFTACSDSFDFGPDHVWTLF

HSYAFDFSVWEIWGALLHGGRLVVVPFEVTRSPAEFLALLAEQQVTLLSQTPSAFHQLTEAARQEPARCAGLALRH

VVFGGEALDPSRLRDWFDLPLGSRPTLVNMYGITETTVHVTVLPLEDRATSLSGSPIGRPLADLQVYVLDERLRPV

PPGTVGEMYVAGAGLARGYLGRPALTAERFVADPNSRSGGRLYRTGDLAKVRPDGGLEYVGRGDRQVKIRGFRIEL

GEIEAALVTHAGVVQAVVLVRDEQTDDQRLVAHVVPALPHRAPTLAELHEHLAATLPAYMVPSAYRTLDELPLTAN

GKLDRAALAGQWQGGTRTRRLPRTPQEEILCELFADVLRLPAAGADDDFFALGGHSLLATRLLSAVRGTLGVELGI

RDLFAAPTPAGLATVLAASGTALPPVTRIDRRPERLPLSFAQRRLWFLSKLEGPSATYNIPVAVRLTGALDVPALR

AALGDVTARHESLRTVFPDDGGEPRQLVLPHAEPPFLTHEVTVGEVAEQAASATGYAFDITSDTPLRATLLRVSPE

EHVLVVVIHHIAGDGWSMGPLVRDLVTAYRARTRGDAPEYTPLPVQYADYALWQHAVAGDEDAPDGRTARRLGYWR

EMLAGLPEEHTLPADRPRPVRSSHRGGRVRFELPAGVHRSLLAVARDRRATLFMVVQAALAGLLSRLGAGDDIPIG

TPVAGRGDEALDDVVGFFVNTLVLRTNLAGDPSFADLVDRVRTADLDAFAHQDVPFERLVEALAPRRSLARHPLFQ

IWYTLTNADQDITGQALNALPGLTGDEYPLGASAAKFDLSFTFTEHRTPDGDAAGLSVLLDYSSDLYDHGTAAALG

HRLTGFFAALAADPTAPLGTVPLLTDDERDRILGDWGSGTHTPLPPRSVAEQIVRRAALDPDAVAVITAEEELSYR

ELERLSGETARLLADRGIGRESLVAVALPRTAGLVTTLLGVLRTGAAYLPLDTGYPAERLAHVLSDARPDLVLTHA

GLAGRLPAGLAPTVLVDEPQPPAAAAPAVPTSPSGDHLAYVIHTSGSTGRPKGVAIAESSLRAFLADAVRRHDLTP

HDRLLAVTTVGFDIAGLELFAPLLAGAAIVLADEDAVRDPASITSLCARHHVTVVQATPSWWRAMLDGAPADAAAR

LEHVRILVGGEPLPADLARVLTATGAAVTNVYGPTEATIWATAAPLTAGDDRTPGIGTPLDNWRVHILDAALGPVP

PGVPGEIHIAGSGLARGYLRRPDLTAERFVANPFAPGERMYRTGDLGRFRPDGTLEHLGRVDDQVKVRGFRIELGD

VEAALARHPDVGRAAAAVRPDHRGQGRLVAYVVPRPGTRGPDAGELRETVRELLPDYMVPSAQVTLTTLPHTPNGK

LDRAALPAPVFGTPAGRAPATREEKILAGLFADILGLPDVGADSGFFDLGGDSVLSIQLVSRARREGLHITVRDVF

EHGTVGALAAAALPAPADDADDTVPGTDVLPSISDDEFEEFELELGLEGEEEQW*

SEQ ID NO: 10
Nucleotides 38555-56047 of SEQ ID NO: 1

SEQ ID NO: 11

VNRRSKVVEEILPVSALQEGLLFHSSFAAADGVDVYAGQLAFDLVGAVDTGRLRAAVE
SLVARHGVLRSSYRQARS

GEWVAVVARRVATPWRAVDARDGATDAAAVAREERWRPFDLGRAPLARFVLVRTDDDRFRFVITYHHVILDGWSLP

VLLRELLALYGSGADPSVLPPVRPYGDFLRWAAARDDAAAETAWRDALTGLDEPSLVAPGASPDGVVPASVHAELD

KAGTENLAAWARHRGITQATAVRAAWALVLGQHTGRDDVVFGVTVSGRPAELAGAEHMVGLFINTVPLRTVLDPAD

TLGTFAARLQAEQTTLLEHQHVRLSDIQRWAGHKELFDTIVVFENYPIGHSGPGSIRTDDFTVTATEGSDATHYPL

TLTAVPGETLRLKLDHRPDLVDTTTATALLRRVTRVLETATDDTGHTLARLDLLDDDERHRLLRGWNDTTREQPPT

YYHQEFEEQARRRPHDTALVFTSTSWTYEELNDRANRLARLLVAAGAGSDDFVALAFPRSAESVVAILAVLKAGAA

YLPLDMDQPAERLTGILADAHPTVVLTTTTATPLPHPGRTLVLDSPTTARALAAAPAHNLTDADRRTPLNARNAAY

IIHTSGSTGRPKGVVIEHRSLANLFHDHRRALIEPHAAGGSRLKAGLTASLSFDTSWEGLICLAAGHELHLIDDDT

RRDAERVAELIDRQRIDVIDVTPSFAQQLVETGILDEGRHHPAAFMLGGEGVDAKLWTRLSDVPGVTSYNYYGPTE

FTVDALACTVGIAPRPVIGHPLDNTAAYILDGFLRPVPEGVAGELYLAGTQLARGYAGRPGLTAERFVACPFGAPG

ERMYRTGDLVRRSPGGVVEYLGRVDDQIKLRGFRIEPAEIELALAGHPAVAQNVVLLHRSATGEARLVAYVVPGTP

VDPRELTGHLAARLPAYMVPSAFVLLDTLPLTPNGKLDRGALPEPAFGTAPRPERPRTPVEEILCGLYADVLGLPS

FGADDDFFDAGGHSLLASKLVSRIRTNLKTELNVRALFEHRTVSSLATALHRAAQAGPALTAGPRPARIPLSYAQR

RLWFLNRLDRDSAAYNMPVALRLRGPLDSTAMCAALTDVAERHEALRTVFEEDRDGAHQIVLPATGLGPLLTVTGA

DGTTLRALITEFVRRPFDLAAEIPFRAALFRVGDEEHVLVVVLHHIAGDGWSMGPLARDVAEAYRARAAGRAPDWE

PLPVQYADYALWQREVLGAEDDETGELSAQLAHWRTRLAGAPAELTLPTDRPRPAVASTAGDRVEFTVPAGLHQAL

ADLARAHGATVFMVVQAALAVLLSRLGAGDDIPIGTPVAGRTDEATEELIGFFVNTLVLRTDVSGDPTFAELLARV

RATDLDAYAHQDVPFERLVEVLNPERSLARHPLFQVMLTFNVPDMDGVGSALGNLGELEVSGEAIRTDQTKVDLAF

TCTEMYAADGAASGMRGVLEYRLDVFGAVQARETTERLVRVLEGVVSGGGGVSVSGVDVLGVGERERLLGWGVGGP

VPVVPGGGLVGLFEERVRADADAVAVRGAGVVWSYGELNARVNVVARWLVGRGVGAECGVGVVMGRGVDVVVMLLA

VAKAGGFYVPVDPEWPVERVGWVLADAGVGLVVVGEGLSHVVGDFPGGEVFEFSRVVRESCLVELVAADGVEVRNV

TDGERASRLLPGHPLYVVYTSGSTGRPKGVVVTHASVGGYLARGRDVYAGAVGGVGFVHSSLAFDLTVTVLFTPLV

SGGCVVLGELDESAQGVGASFVKVTPSHLGLLGELEGVVAGNGMLLVGGEALSGGALREWRERNPGVVVVNAYGPT

ELTVNCAEFLIAPGEEVPDGPVPIGRPFAGQRMFVLDAALRVVPVGVVGELYVAGVGLARGYLGRAGLTAERFVAC

PFGAPGERMYRTGDLVRWRVDGALEFVGRADDQVKVRGFRVELGEVEGAVAAHPDVVRAVVVVREDRPGDHRLVAY

VTGVDTGGLSSAVMRAVAERLPAYMVPSAVVVLDEIPLTPNGKVDRAALPVPGVEAGAGYRAPVSPREEVLCGLFA

EVLGLERVGVDDDFFGLGGHSLLATRLISRVRAVLGVEAGVRALFEAPTVSRLERLLRERSALGVRVPLVARERTG

REPLSFAQQRLWFLEELEGPGAAYNIPMALRLAGVLDVEALHQALIDVIARHESLRTLIAQDAGTAWQHILPVDDP

RTRPGLPLVDIGADALQERLDEAAGRPFDLAADLPVRATVFRLTDNDHILLVVAHHVAFDAMSRVPFIRNVKRAFE

ARTNGAAPDWRPLPVQYADYAAWQRDVLGTEDDESSELSAQLAYWRTQLASLPAELALPTDRARPAVASYEGGKVE

FTVPAGVYDGLVALARAEGVTVFMVVQAALAALLSRLGAGDDIPIGTPIAGRTDQATEDLIGFFVNTLVLRTDVSG

DPTFAELLARVRATDLDAYAHQDIPFERLVEAVNPERSLARHPLFQVMLTFDNTIDREVTEGFAGLGVEGLPLGAG

AVKFDLLFGLSEVGGELRGAVEYRCDLFDHPTVAQLAERLVRVLERVASDASVRTGELPVVGEAERARVLTEWNDT

GVPGVPETFLELFEAQVAARGDAPAVVYEGEVLSYRELDARANRLAGLLVGRGAGPEHFVGVALPRGLDLIVALLA

VLKSGAAYVPLDPEYPAERLVHMVTDAAPVVVVTSTDVRTLRTVPRVELDDEATRATLVAAPATGPDVKMSASHPA

YVIYTSGSTGRPKGVVISHGSLANFLAWAREDLGAERLRHVVLSTSLSFDVSVVELFAPLSCGGTVEIVRNLLALV

DRPGRWSASLVSGVPSAFAQLLEAGLDRADVGMIALAGEALSARDVRRVRAVLPGARVANFYGPTEATVYATAWYG

DTPMDAAAPMGRPLRNTCVYVLDDGLRVVPVGVVGELYVAGVGLARGYLGRVGLTAERFVACPFGARGERMYRTGD

LVRWRVDGTLEFVGRADDQVKVRGFRVELGEVEGAVAAHPDVVRAVVVVREDRPGDHRLVAYVTGVDTGGLSSAVM

RAVAERLPAYMVPSAVVVLDEIPLTPNGKVDRAGLPVPVVSVAGFCAPSSPREEVLCGLFAEVLGVERVGVDDGFF

DLGGDSILSIQLVARARRAGLELSVRDVFEGRTVRALAAVVRGSDAGAVGVVGGAEIVLPGVGEVERWPVVEWLAE

RGGGSLGGVVRGFNQSVVLAVPAGLVWEELRVLLGAVRDRHEAWRLRVLDSGALCVDGVVPDDGSWIVRCDLSGMG

VDGQVDAVRAAAVEARAWLDPSVGRVVRAVWLERGGDRSGVLVLVAHHLVVDGVSWRVVLGDLAEGWAQVRSGGRV

ELGVVGTSLRGWAAALAEQGRRGERAGEVELWSRMVRGADVLVGSRAVDGAVDVFGGVVSVDSRASVSVSRALLTE

VPSVLGVGVQEVLLAAFGLAVARWRGRGGPVVVDVEGHGRNEDAVRGADLSRTVGWFTSVYPVRVPVESASWDEVR

AGGPVVGRVVREVKETLRSLPDQGLGYGILRYLDPEHGPALARHATPQFGFNYLGRFTTGTDDTGDEGMTDWVPVS

GPFAVGAGQDPELPVAHAVEFNAITLDTPEGPRLGVTWSWPTTLLPESRIRELARYWDEALEGLVEHARHPEAGGL

TPSDVTLVEVNQVELDRLQAGVAGGAEEILPVSALQEGLLFHSALASGGVDVYVGQLVFDLVGPVDVDRLRAAVEG

LVARHGVLRSGYRQLRSGEWVAVVARQVDLPWQSIDVRDGGIDGLVEEERWRRFDMGRGPLARFVLIRTHDDRFRF

VITYHHVVLDGWSVPVLLRELLALYGSSGDVSVLPGVRSYGDFLRWVAARDAAAAEGAWRRALTGLEEPSLVAPGV

SRDGVVPAAFHGAVDGDLSQKIVAWARGRGVTVASVVQAAWALVLGRLMGRDDVVFGVTVSGRPAEVVGVEDMVGL

FVNTIPLRARLDPAESLGGFVERLQREQTELLEHQHVRLAEVQRWAGHKELFDVGMVFDNYPVSSESPEAEFQISR

TGGYNGTHYALNLVASMHGLELELEIGYRPDVFDAGRVREVWGWLVRVLEGVVSGGGGVSVSGVDVLGVGERERLL

G*

SEQ ID NO: 12
Nucleotides 56044-68361 of SEQ ID NO: 1

SEQ ID NO: 13

VRGVGGPVPVVPGGGLVGLFEERVRADADAVAVRGAGVVWSYGELNARVNVVARW
LVGRGVGAECGVGVVMGRGVD

VVVMLLAVAKAGGFYVPVDPEWPVERVGWVLADAGVGLVVVGEGLSHVVGDFPGGEVFEFSRVVRESCLVELVAAD

GVEVRNVTDGERASRLLPGHPLYVVYTSGSTGRPKGVVVTHASVGGYLARGRDVYAGAVGGVGFVHSSLAFDLTVT

VLFTPLVSGGCVVLGELDESAQGVGASFVKVTPSHLGLLGELEGVVAGNGMLLVGGEALSGGALREWRERNPGVVV

VNAYGPTELTVNCAEFLIAPGEEVPDGPVPIGRPFAGQRMFVLDAALRVVPVGVVGELYVAGVGLARGYLGRVGLT

AERFVACPFGVPGERMYRTGDLVRWRVDGALEFVGRADDQVKVRGFRVELGEVEGAVAAHPDVVRAVVVVREDRPG

DHRLVAYVTAGGVGGDGLRSAISGLVAERLPAYMVPSAVVVLDEIPLTPNGKVDRAALPVPEVEAGTGYRAPVSPR

EEVLCGLFAEVLGVERVGVDDDFFELGGHSLLATRLISRVRAVLGVEAGVRALFEAPTVSRLERLLRERSGLGVRV

PLVARERTGREPLSFAQQRLWFLEELEGPGAAYNIPMALRLAGVLDVEALHQALIDVIARHESLRTLIAQDAGTAW

QHILPVDDPRTRPGLPLVDIGADALQERLDEAAGRPFDLAADLPVRATVFRLTDNDHILLLVLHHIAGDGWSMGPL

ARDLSTAYSARAAGAASAWRPLSVQYADYAAWQRDVLGTEDDESSELSAQLAYWRTQLASLPAELALPTDRARPAV

ATYRGGRIEFTIPADVHRSLADLARAEGVTVFMVVQAALAALLSRLGAGDDIPIGTPIAGRTDQATEDLIGFFVNT

LVLRTDVSGDPTFAELLARVRATDLDAYAHQDIPFERLVEAVNPERSLARHPLFQVMLAFNNAETSTPLPMAEGLA

ASRQDIEPGVAKFDLALYCNESRGETGDHQGIRSVFEYRRDLWDEDTVRQLADRFLHVLAAFAAAPEQRASSVDVL

RAGERDQLLHEWNDTAAALPPALLPQLFEEQVRRTPHDVALVSGNIRLTYAELDARANRLAHLLLARGAAPETFVA

VALPRTEELLVALLAVQKTGAGHLPLDPGFPAERLSYMLDDARPAVVLTTEDISARIPGGSHVVLDSEQVTGELHD

HPATSPAGRGNPAGPAYVIYTSGSTGQPKGVVVPSAALVNFLADMVPRLGLRGGDRLLSVTTVGFDIAALELFVPL

LSGATVVLADGETVRDPALARQTCEDHGVTMVQATPSWWHGMLADAGDSLRGVHAVVGGEALSPGLRDALTRGARS

VTNMYGPTETTIWSTSAGQAAGDSAPPSIGTPILNTRVYVLDAALCVVPPGVAGELYIAGDGLARGYLGRAGLTAE

RFVACPFGAPGERMYRTGDLVRWRVDGALEFVGRADDQVKVRGFRVELGEVEGAVAAHPDVVRAVVVVREDRPGDH

RLVAYVTGVDTGGLSSAVMRAVAERLPAYMVPSAVVVLDEIPLTPNGKVDRAALPVPGVEAGAGYRAPVSPREEVL

CGLFAEVLGVERVGVDDDFFGLGGHSLLATRLISRVRAVLGVEAGVRALFEAPTVSRLERLLRERSGLGVRVPLVA

RERTGREPLSFAQQRLWFLEELEGPGAAYNIPMALRLAGVLDVEALHQALIDVIARHESLRTLIARDSDGTARQQV
LPVGDPAARPALPVVQTDADTLVAKLNEAVGRPFDLTAEMPLRATVFRVADEDHALLLVFHHIAGDGWSTGLLARD
LSTAYAARLEGRDPQLPPLPVQYADYAAWQRDVLGTEDDESSELSAQLAYWRTQLADLPAELALPADRVRPARASY
EGGRVGFTVPAGVLRDLTRLARVEGVTVFMVVQAALAALLSRLGAGDDIPIGTPIAGRTDQATEDLIGFFVNTLVL
RTDVSGDPTFAELLARVRATDLDAYAHQDIPFERLVEAVNPERSLARHPLFQVMLAFDNTADGGPVEDFPGLSAAG
LPLGAGAAKFDLLFGLSEVGGELRGAVEYRCDLFDHPTAARIAERLVRVLERVAADASVRLGELPVVSDAERACVL
TEWNDTAVPGVTGTLSALFEARAAARGDAPAVVYEGEELSYRELNTRANRLAHVLAEHGAGPERFVGVALPRSPDL
VVALLAVVKSGAAYVPLDPEYPADRLAYMAGDAAPVAVLTRGDVELPGSVPRIGLDDTEIRATLATAPGTNPGTPV
TEAHPAYMIYTSGSTGRPKGVVVSHGAIVNRLAWMQAEYRLDATDRVLQKTPAGFDVSVWEFFWPLLEGAVLVFAR
PGGHRDAAYLAGLIERERITTAHFVPSMLRVFLEEPGAALCTGLRRVICSGEALGTDLAVDFRAKLPVPLHNLYGP
TEAAVDVTHHAYEPATGTATVPIGRPIWNIRTYVLDAALRPVPPGVPGELYLAGAGLARGYHGRPALTAERFVACP
FGVPGERMYRTGDLVRWRVDGTLEFVGRADDQVKVRGFRVELGEVEGAVAAHPDVVRAVVVVREDRPGDHRLVAYV
TVGGVGGDGLRSAISGLVAERLPAYMVPSAVVVLDEIPLTPNGKVDRAGLPVPVVSVAGFCAPSSPREEVLCGLFA
EVLGVERVGVDDGFFDLGGDSILSIQLVARARRAGLELSVRDVFEGRTVRALAAVVRGSDAGAVGVVGGAEIVLPG
VGEVERWPVVEWLAERGGGSLGGVVRGFNQSVVLAVPAGLVWEELRVLLGAVRDRHEAWRLRVLDSGALCVDGVVP
DDGSWIVRCDLSGMGVDGQVDAVRAAAVEARAWLDPSVGRVVRAVWLERGGDRSGVLVLVAHHLVVDGVSWRVVLG
DLAEGWAQVRSGGRVELGVVGTSLRGWAAALAEQGRRGERAGEVELWSRMVRGADVLVGSRAVDGAVDVFGGVVSV
DSRASVSVSRALLTEVPSVLGVGVQEVLLAAFGLAVARWRGRGGPVVVDVEGHGRNEDAVRGADLSRTVGWFTSVY
PVRVPVESASWDEVRAGGPVVGRVVREVKETLRSLPDQGLGYGILRYLDPEHGPALARHATPQFGFNYLGRFTTGT
DETTTADALDRAPAWSLLARSAAGQDPELPVAHAVEFNAITLDTPEGPRLGVTWSWPTTLLPESRIRELARYWDEA
LEGLVEHARHPEAGGLTPSDVGLAELSFAEIELLEDDWRTQG*

SEQ ID NO: 14
Nucleotides 68358-78062 of SEQ ID NO: 1

SEQ ID NO: 15
VSESRCAGQGLVGALRTWARTRARETAVVLVRDTGTTDDTASVDYGQLDEWARSIAVTLRQQL
APGGRALLLLPSGPEFTAAYLGCLYAGLAAVPAPLPGGRHFERRRVAAIAADSGAGVVLTVAG
ETASVHDWLTETTAPATRVVAVDDRAALGDPAQWDDPGVAPDDVALIQYTSGSTGNPKGVVVT
HANLLANARNLAEACELTAATPMGGWLPMYHDMGLLGTLTPALYLGTTCVLMSSTAFIKRPHL
WLRTIDRFGLVWSSAPDFAYDMCLKRVTDEQIAGLDLSRWRWAGNGAEPIRAATVRAFGERFA
RYGLRPEALTAGYGLAEATLFVSRSQGLHTARVATAALERHEFRLAVPGEAAREIVSCGPVGH
FRARIVEPGGHRVLPPGQVGELVLQGAAVCAGYWQAKEETEQTFGLTLDGEDGHWLRTGDLAA
LHEGNLHITGRCKEALVIRGRNLYPQDIEHELRLQHPELESVGAAFTVPAAPGTPGLMVVHEV

RTPVPADDHPALVSALRGTINREFGLDAQGIALVSRGTVLRTTSGKVRRGAMRDLCLRGELNI
VHADKGWHAIAGTAGEDIAPTDHAPHPHPA*

SEQ ID NO: 16
Nucleotides 36,408-38,201 of SEQ ID NO: 1

SEQ ID NO: 17
MNPPEAVSTPSEVTAWITGQIAEFVNETPDRIAGDAPLTDHGLDSVSGVALCAQVEDRYGIEV
DPELLWSVPTLNEFVQALMPQLADRT*

SEQ ID NO: 18
Nucleotides 38,270-38,539 of SEQ ID NO: 1

SEQ ID NO: 19

MIGVAPPAYDPAAPESATTLPVGTPTTVRSYVRSLLRRHRRAFTVLIAVNAVAVVASITGPYL
LGGLVEDLSAGVTDLHLERTAAIFAVALVVQVLFTRSMRLRGAMLGEEMLADLREDFLVRSVG
LPPGVLERAGTGDLLSRITTDIDRLANAMREAVPQLAIGVVWAGLLLGALTVTAPPLALAVLI
ALPVLIVGCRWYFRRAPSAYRSEAAGYAAVAAMLAETVDAGRTVEAHRLGGRRVALSDRRISQ
WTAWERYTLFLRSVLFPVINATYVTILGAVLLLGGWFVLEGWLTVGQLTTGALLAQMMVDPIG
LILRWYDELQVAQVSLARLVGVRDIEPDAGDAEVGPEGRDVRADEVRFGYREGVDVLHKVSLD
VAPGTRLALVGPSGAGKSTLGRLLAGIYAPRTGEVTLGGAELSRMTAERVREHVALVNQEHHV
FVGSLRDNLRLAREGAKDAELWASLAAVDADGWAKALEKGLDTEVGSGGFTLTPAQAQQIALA
RLVLADPHTLVLDEATSLLDPRAARHLERSLARVLEGRTVV*

SEQ ID NO: 20
Nucleotides 1637-1 of SEQ ID NO: 1

SEQ ID NO: 21
MSPPAPPEALQRPAPTAQEPVRTGSRTGLVAICVSLFAALVVSVVVAIGLGPAVVPPAETARF
LWAALSGGPISADEVTTYQIIWQIRTPRVLLAALVGAGLSAVGVAIQALVRNALADPFVLGVS
SGASVGAVGVTVMGGLAVFGIYAVSVGAFLGALVASVLVYGASSTKGALSPLRLVLTGVAMSL
GFQAVMSVIIYFAPSSEATSMVLYWTMGSFGAASWGSLPVVTAAVLLGVLVLHRHGRPLDVLA
LGDETAASLGISPDRHRKSLLVLVSLVTGVMVAVSGSIAFVGLVMPHLVRMVVGATHARVLAV
APLAGAVFMVWVDLVSRTLVAPRELPLGVITALVGVPVFITLMRRKSYMFGGR*

SEQ ID NO: 22
Nucleotides 3502-1634 of SEQ ID NO: 1

SEQ ID NO: 23
MNDDARPAPEPQDIPPHSGAADEVNRQDPSRRSVLWTTAGVAGAGLGLGALGAGTASAAGRSAPDAVAAAEAVAAA
PPRQGRTMAGVPFERRSTVRVGIIGLGNRGDSMIDLFLALPGVQVKAVCDTVRDKAEKAAKKVTAAGQPAPAIYAK
DEHDYENLCKRGDIDFVYVVTPWELHFPMAKTAMLNGKHVGVECPIAMRLEELWQLVDLSERTRRHCMQLENCCYG
KNEMRVLRMAHAGLFGELQHGAGAYNHDLRELMFDPDYYEGPWRRLWHTRLRGDLYPNHGFGPVANYMDVNRGDRV
VSISSVGTTPLGLAAYREEHMPAGDPSWKESYIGADRTISLVQTAKGRVIRLEHDVSSPHPYSRINSLGGTKGVEE
DYPERIYLEPTNTNHQWDDFKKYAEWDHWLWKEHANPPGGHGGMDYIMVFRLMQCMRLGLVPDFDVYDAAVWTAPV
PLSHLSIKAKGVPLPIPDFTRGEWKKTRSGMDSEKPAE*

SEQ ID NO: 24
Nucleotides 4927-3659 of SEQ ID NO: 1

SEQ ID NO: 25

MPLLEPDPEALRPGTAREPAPDRVTDGSAGGTPEPLRSELTALLGADKVLWKISDLVRYASDASPYRFLPRVVLVP
EDLDDVSAILSYAHGKGRSVVFRAAGTSLNGQAQGEDILVDVRRHWTGVEVLDDGARARILPGTTVMRANAALARY
GRLLGPDPASAIACTLGGVVANNASGMTAGTTRNSYRTLASLTFVLPSGTVVDTAHPAADEELAHAEPELCAGLLE
LKAEIEADAELTARIRAKYTIKNTNGYRLDAFLDGATPVQILRGLMVGSEGTFGFISEVVFDTLPLDRRVSSGLLF
FPSLTAAAAAVPRFNEAGAIAVELMDGNTLRASVSVPGVPADWAALPRETTALLVEFRAADEAGRAAFERAADAVV
AGLDLVRPAASVTNAFTRDAGTIAGYWKARKAFVTAVGGSRPSGTTLITEDFAVPPARLADACAALLELQSRHGFD
AAVAGHAAHGNLHFLLAFDAAKPADVARYDAFMQEFCALVVDRFDGSLKAEHATGRNIAPFLEREWGPRATELMWR
TKQVIDPAGVLAPRIVLDRDPRAHLRGLKTIPKVEAVADPCIECGFCEPTCPSEDLTTTPRQRIVLRREMMRQTDG
SPVESGLLDAYGYDAVDTCAGDSTCKLACPVGIDTGAMMKGFRHRRHTPREERIAALTAKNFRAVEASARLAVAAA
DTVGNRVGDAPLQAVTRLARKAVRPDLVPEWLPQIPGAAARRLPDTARVGASAVYYPACVNRIFAGPDDGDAGPAL
SLAEAVVAVSGRAGKPVWIPEDVTGTCCATIWHSKGYDAGNRIMANRIVEAAWGWTAGGTLPLVVDASSCTLGIAE
EVVPYLTEDNRALHRELTVVDSLVWAAEELLPHLTVFRTAGSAVLHPTCSMEHLGDVGQLRALAEACAQEVVVPDD
AGCCAFAGDRGMLHKELTDSATAKEAAEVDRRPYDAYLSANRMCEIGMERATGHPYRSALIELEHATRPTLP*

SEQ ID NO: 26
Nucleotides 8364-5410 of SEQ ID NO: 1

SEQ ID NO: 27
MDAPDSPDSPDSPESRDSRDSRDSRDGLLAEQLLRLTRRLHRIQRRQLEPIDITPAQFRLLRTVASYDAAPRMADL
ARRLDVVPRAVTTLVDALEASGRVRRAPDPDSRRVVRIEITDEGRATLRSLRSARRAAAEEILAPLTADQREVFGE
LLSALVDGMPERHC*

SEQ ID NO: 28
Nucleotides 8916-8416 of SEQ ID NO: 1

SEQ ID NO: 29
MKPDEPTWTPPPDARPAADRRPAEVRRILRLFRPYRGRLAVVGLLVGASSLVGVASPFLLREILDTAIPQGRTGLL
TLLALGMILTAVMTSVFGVLQTLISTTVGQRVMHDLRTAVYTQLQRMPLAFFTRTRTGEVQSRIANDIGGMQATVT
STATSLVSNLTAVIATVVAMLALDWRLTVVSLLLLPVFVAISRRVGRERKKITTQRQKQMAAMAATVTESLSVSGI
LLGRTMGRSDSLTQGFAEESERLVDLEVRSNMAGRWRMSVIGIVMAAMPAVIYWAAGLTFASGAAAVSIGTLVAFV
TLQQGLFRPAVSLLSTGVQMQTSLALFQRIFEYLDLTVDITEPEHPVRLERIRGEIAFEDVDFSYDEKNGPTLTGI
DVTVPAGDSLAVVGSTGSGKSTLSYLVPRLYDVTGGRVTLDGIDVRDLDFDTLARAVGVVSQETYLFHASVADNLR
FAKPEATDEEIEAAARAAQIHDHIASLPDGYDTMVGERGYRFSGGEKQRLAIARTILRDPPVLILDEATSALDTRT
EQAVQEAIDALSAGRTTLTIAHRLSTVRDADQIVVLEDGRVAERGTHEELLDRDGRYAALIRRDSHPVPVPVPAP*
*

SEQ ID NO: 30
Nucleotides 9030-10853 of SEQ ID NO: 1

SEQ ID NO: 31
HRHLAERPRRCAVLALLRPAAGPAGRAGRRPGPAARSDPLHRQGGRRPHRDIGEAAGRAARPAADTQTAAAEPAQR
PGVHRQLHRAARRMQHRGEDPGGGARHDGHAGSRGDGQARPRPVLPAAPLRPRGPGRAALSHGGSRPVGRGVPGPS
AHPGPPDARHPRGGGDGGTRVRRAAALHRTGSGERLSRPAAYTQHTAHRAHGAHSTHGGAAAPVGRGATAPGGAMV
RRANPRSGRRRQAGWSGSSSGLSPCTWCICGTAQ*

SEQ ID NO: 32
Nucleotides 10933-11544 of SEQ ID NO: 1

SEQ ID NO: 33
MVNESPDARPRRRLRPTRRGKIVLVVGALLVVTAAVLIPLSLTGSDEPPKKQETPQSTLMIPEGRRVSQVYEAVDK
ALDLKPGSTLKAASTVDLKLPAQAEGNPEGYLFPATYPIDDTTEPAGLLRYMADTARKHFAADHVTAGAQRNNVSV
YDTVTIASIVQAEADTPADMGKVARVVYNRLLKDMPLQMDSTINYALKRSTLDTSTADTQLDSPYNSYRIKGLPPT
PIGNPGEDALRAAVRPTPGPWLYFVTVGPGDTRFTDSYDEQQKNVEEFNRGRGSATTG*

SEQ ID NO: 34
Nucleotides 11990-12850 of SEQ ID NO: 1

SEQ ID NO: 35
MIPGARRVSRSVNISGVRELDVVVIGAGQAGLSAAYHLRRVGLEPDNDFVVLDHAPRPGGAWQFRWPSLTYGKVHG
MHALPGMELTGADPDRPSSEVIGAYFAAYEDRFGLRVHRPVEVSAVREGSGGRLLVETSEGTYAARALINATGTWD
RPFWPRYPGQETFRGRQLHTANYPGPEEFAGQRVLVVGGGASGTQHLMEIAEHAADTFWVTRSEPVFREGPFTEEW
GRAAVAMVEERVRNGLPPKSVVSVTGLPLNDAVRRARERGVLDRLPMFDRITPTGVAWDDGRTVETDVILWATGFR
PAVDHLAPLKLREPGGGIRAEDTRAVRDGRVHLVGYGPSASTIGANRAGRAAVRSVMRLLKETGADGGASAVVSVP
APVPGV*

SEQ ID NO: 36
Nucleotides 14038-12878 of SEQ ID NO: 1

SEQ ID NO: 37
VPGLARPTRSTPPRQLRRGHPPSLSRPPTEPLTTPPPPEPPTQRHTSLCNTDSLAVAMSERPRHRPQKRSIACGAC
RAGSSPLAHTGVGLVRGGAGTALVGSHAEVADRIEEYHALGVEHFVLSGYPHLEEAYWFGEGVTPELSRRGLLSTV
PASPLLGVSGAESRTATAPGGAPLLLAGGR*

SEQ ID NO: 38
Nucleotides 14348-14070 of SEQ ID NO: 1

SEQ ID NO: 39
VAVVAEDLRRRFAATKVTFLIVDLTGRALARLSTTTAAGSENETERIPLFGGSVYEQVIRTQRPHHEPAGQEQRVI
VPVTNRGDAIGLLELLLPAGRSDEEEVVLAVGEAAHALAYVVIANGRFTDFYTWGKRSRPPTLAAEIQYQLLPQAL
SCEAAQFTLSGSLEPSEDLSGDTFDYALDRDTLHLSVTDPMGHDLGAALAATVLVGALRRARRAGAPLAEQARQGD
QALTSHGQGHATGQLLRINLHTGKAELVNAGHPWPMRMRAGMVETIPCQVDQPFGLAVVSPRPYRVQTLDLHPGDR
LLMLTDGMLERHGEKIDVAALLRQTRSLHPRETTLMLTSAVRDAAGGRLEDDATVVCLDWHGPQEVHRHVSSGADT
HQASAARPPNR*

SEQ ID NO: 40
Nucleotides 15697-14522 of SEQ ID NO: 1

SEQ ID NO: 41
MRVRLQVGVALCGLGVLVTQERERRRCGARSAGMVPDPLLLAVAFEAGAFAFQGASRSRVRAEHGQGGALRQTARK
FANSGPATGRAVGQDDPMSQDLVTFLHARLDEEANLAGRCDGDGCGEWAPHGHTVDFCQGELSGFHSTIALHVALH
DPARVLREAEAKRRVLARHGLSPATGDPELPWDNRDDCRYDGATWPCDDLLGLASPYADHPDYPQRP*

SEQ ID NO: 42
Nucleotides 17597-16938 of SEQ ID NO: 1

SEQ ID NO: 43
MSVRDLVGMPCHPCEPPRRAEGRRRGVGRMRWWKGVLMTVRHQGVRWWFALLALVGCVVCVLCVVALSGAGHYFGL
SLWAGIALVVVGALFPLGGLGFLYWVDDGRSEDSFLVKFLCFVAHSAVLGLAAVSCTGAEAWAFEQRGRWTEATVV
GYSPPRVVPGDPPTKVRASCALETAEGERVRPRLPEGRGCRDGVRHGSRLDVLYDPRGLLAPRATEPMDHGVTVPV
LGGVATLSGFLGCVALAWRWETLRVRSARRTAARRGRESAAG*

SEQ ID NO: 44
Nucleotides 17870-18682 of SEQ ID NO: 1

SEQ ID NO: 45
MKFTKLAIPVAASALLLTGCGAEVESQGKGSGKSTVKRCGESVEYTVPKRAVAYEGGSADKLFSLGLADHVHGYVM
PPANPPVSESPWAKDYAKVKMLSDDLLNKEIVVDAKSDFVVAGWNSGFSDQRGITPEILDKLGVQSFMHSESCYNY
PGHPEKLTPFKGLYTDLERLGRIFQVEEEAEKVVAGLKKREAAVAEQAPKGKPVPVFLYDSGTDQPFTAGNQVPPN
DIIKTAGGKNIFDGLEERWTQVNWEAVTQAEPEVIMIFDYGDQPAEKKIEFLKKSPHTKELPAVKKNNFFVLDYNE
GISSPRNIDGLEKFGKYMRAFKK*

SEQ ID NO: 46
Nucleotides 19898-18915 of SEQ ID NO: 1

SEQ ID NO: 47
MDLELDGLSVVTDGKSLVRDLSLDVGSGQVVGLVGPNGSGKSTALRCVYRALKPSSGTVKVDGQELSSLTMRRSAQ
LIAAMTQDGAVDLDFTVEEVIALGRTPHQRGSTPLNGHERDLCEHAMRRLDILHLARRGILTLSGGERQRVLLARA
LVQEPKILVLDEPTNHLDVRHQVRLLSLLRGAGLTVLVVLHDLNLAAAACDRIGVLSEGRLITSGTPKDVLTPELV
DEVFGVRASVVPHPLTGDPQLLYSLDS*

SEQ ID NO: 48
Nucleotides 20674-19907 of SEQ ID NO: 1

SEQ ID NO: 49
MSPPAPPEALQRPAPTAQEPVRTGSRTGLVAICVSLFAALVVSVVVAIGLGPAVVPPAETARFLWAALSGGPISAD
EVTTYQIIWQIRTPRVLLAALVGAGLSAVGVAIQALVRNALADPFVLGVSSGASVGAVGVTVMGGLAVFGIYAVSV
GAFLGALVASVLVYGASSTKGALSPLRLVLTGVAMSLGFQAVMSVIIYFAPSSEATSMVLYWTMGSFGAASWGSLP
VVTAAVLLGVLVLHRHGRPLDVLALGDETAASLGISPDRHRKSLLVLVSLVTGVMVAVSGSIAFVGLVMPHLVRMV
VGATHARVLAVAPLAGAVFMVWVDLVSRTLVAPRELPLGVITALVGVPVFITLMRRKSYMFGGR*

SEQ ID NO: 50
Nucleotides 21782-20676 of SEQ ID NO: 1

SEQ ID NO: 51
VSAGTSRSAVAPEKSPEMPGDLKMARALWPVLVASAVGLLPFTVFSTYLVPIAEETGSGVAAVGGLRGLGGLAALA
VGTALAPLIDRVPKSKAVAVGLVVLAVSSALGASGDFLLTAVFCLLVGAGTAVINPALTAAAADRFGDGKSAARAA
TLVTSTTSMTAMLAAPLIALPALLWGWEGDLLAVTVVSLLLAAVFLVRGRKGEDPVVEGGPRTGYFASFKALAQVR
GSVPLLAISFLRTAVFMGYLAYLAVYYDDRFHLDPALFSLVWTLSGASFFVSNLLTGRITNAEKSTVGTEQLLLVG
LLAALVTATGFWFTTWLPLALAFTSLHAASHAAVAACAVSLLVRRCGSMRGSALSLNAAGQSLGVFAGAALGGAGL
GLAGYPGIAAAFGLLVAVAVVAGLLVLRSEDEIPGSA*

SEQ ID NO: 52
Nucleotides 23130-21877 of SEQ ID NO: 1

SEQ ID NO: 53
MTPPPTRRKPSDMPFPTPQSVAELTDAVLAGDYGPDPKDMTVTSAFWLYHTTRLAGGPVTYHNHYLVLRVGRSFGG
CSFEAGELTPDFCENASGHPLEKLLRHESAPVRIAALDAYLAQIQPHREAPEQEAVPLPVGTPEVRAKARDASIAG
LLDIEEGAKVALIGVVNPLVAAIRERGGVCLPCDLNLRTTQWGEPVADDMTEVLAEAHAVVATGMTLSNGTFDLIL
EHCREQKVPLVVYAQTGSAVARAFLGSGVTALNAEPFPFSQFSADETTMYRYRAGGDL*

SEQ ID NO: 54
Nucleotides 23951-23127 of SEQ ID NO: 1

SEQ ID NO: 55
MYEHIAEAIKKPDLIALRPDLVCLRFETMKIYSALGAVRHLLESGTVKPGDTLVDSSSGIYAQALALACHRYGMKC
HIVGSTTVDRTLKAQLEILGATLEQVRPSRNLRLDQELRVRRIAEILEENPSYHWMRQYHDSIHYYGYREVAETIA
DEVPAGPLTLVGGVGSGASTGAIASYLREAGRDVSLVGVQPFGSVTFGSEHVSDPDMIIAGIGSAIPFENVRHDLY
DRIHWVSFDSALAGAVHLLRSSGIFAGLSAGAAYLTTRWERSKDDSRTYVFIAADTGHRYVDSAYAKHTEAPDIED
LEPREITSLDELSHPWSAMTWTDDSTSDQKKAL*

SEQ ID NO: 56
Nucleotides 24966-23953 of SEQ ID NO: 1

SEQ ID NO: 57
MDTGVGTAYGTFGELLQGELPEEAGDFLVTLPVARWARASFRCDPAMGDVIVRPSHKEKARRLACLILEEAPGMTG
GVLTVNSVIPEGKGLASSSADLVATARAVGRALRLDMPPSRIEGLLRLIEPTDGVLYPGIVAFHHRAVRLRAMLGS
LPAMSVVGVDEGGAVDTVDFNRIPKPFTPADRREYADLLNRLSGAVRSRDLAEVGRVATRSALMNQPLRYKRLLEP
MREICRDAGGLGVAVGHSGTALGVLLDAADPAYPHRATAVARACGDLAGAVAVYRTLSFPNAVSHGGRTVG*

SEQ ID NO: 58

Nucleotides 25228-26127 of SEQ ID NO: 1

SEQ ID NO: 59
MLTAQQPAPGVVPARIHVTDRLEAAHPLAADGAVVLTGVEPSGDGLVLAAAAVLGERLQQVFPHRLRASDGSNFVH
LHADSFDFVVNVGGVEHRRRDPDEDYVLIQCVRQSDSGGDSFVADAYRFVDHCATADPELWDFLTRGDVDLYGAWS
GLRGMPATPFVGRHVEYTRAGRRIVRRGDGVTPLHRDPGADHTRRMLARLEEAVHALEETLPRFRLDKGEILVLDN
YRCWHGREAHTGDRAVRILTVRSSDAR*

SEQ ID NO: 60
Nucleotides 26445-27212 of SEQ ID NO: 1

SEQ ID NO: 61
MTTMFNNNPPFPPATELRNERVRFQRLSAGYPGRPVLHQLSAAIPPLAMTALVGPNGSGKSTLLGVLAGVITATSG
QLRYAEGSPPAFVPQRGAVGDTLPLTARQTVEMGRWGQRGLWRRLTRTDRTAVDSAMERLGVADLGARQLGELSGG
QRQRVLIAQGLAQQSDLLLLDEPTTGLDPEARERITALLTDLVADGTTVVQATHDLDAARSADACLLLADGRLIGQ
GSPEEVLTPEALARIWQPA*

SEQ ID NO: 62
Nucleotides 28124-27381 of SEQ ID NO: 1

SEQ ID NO: 63
MEWLTAPFEVAFVQRALWAGILVSAICALAGTWVVLRGMAFLGDAMSHGLLPGVAVASLLGGNLLVGAVVSAAVMA
AGVTALGRTPRLSQDTGIGLLFVGMLSLGVIIVSRSQSFAVDLTGFLFGDVLAVRGSDLLLLGVALLLALAVSVLG
YRAFLALAFDERKARTLGLRPRLAHAVLLGLLALAIVASFHIVGTLLVLGLLIAPPAAAMPWARSVQAVMVLAALL
GAAATFGGLLLSWHLRTAAGATVSALAVALFFLSHLASGLRHRRRARRGGLAEPAVAPGRDLLHVLTERNLRRSPC
SSEKTSHRWLRRLRP*

SEQ ID NO: 64
Nucleotides 28139-29098 of SEQ ID NO: 1

SEQ ID NO: 65
VILLTAGCGGGDEAKSGSGPASSSPTPHGYVEGATEAAEQQSRLLLGDPGSGETRVLDLITGKVYDIARSPGATAL
TTDGRFGYFHGPDGIRVLDSGAWMVDHGDHVHYYRAKIKEVGELPGGTGTSIRGDAGVTVASSADGKASVYRRADL
EKGALGTPSPLPGTFAGAVVPYAEHLVTLTAESGAPAKVAVLDRSGKRVAAPEAECEEPQGDAVTRRGVVLGCADG
ALLVHEDDGAFTAEKIPYGEDVPKTERAVEFRHRPGSSTLTAPAGKDAVWVLDAGEGAWTRVKTGPVVAANTAGEG
SPLVVLETDGALHGYDIPTGKETGVTDPLLKELPGTGAGGGAAPVIEVDRSRAYLNDPEGKRVYEIDYNDDLRVAR
TFDVDVRPSLMVETGR*

SEQ ID NO: 66
Nucleotides 29095-30285 of SEQ ID NO: 1

SEQ ID NO: 67
MSARVGAPRMRALLVSLAGFFVVAGAATGCAGGGDERPRVVVTTNILGDITREIVGDEAGVSVLMKPNADPHSFGL
SAVQAAELENADLVVYNGLGLEENVLRHVEAARESGVAAFAAGEAADPLTFHAGQDGGPEEDAGKPDPHFWTDPDR
VREAAGLIADQVAEHVEGVDEKKVRENAERYDGQLADLTGWMEKSFAAIPEDRRALVTNHHVFGYLADRFGLRVIG
AVIPSGTTLASPSSSDLRSLTQAMEKAKVRTVFADSSQPTRLAEVLRQEMGGDVDVVSLYSESLTEKGKGAGTYLE
MMRANTSAMAEGLTGD*

SEQ ID NO: 68
Nucleotides 30282-31244 of SEQ ID NO: 1

SEQ ID NO: 69
MNKPTRARVFTGTALVVAASMALTACGGNGNDDAPSGKEPKEQKSSEAAAVGNPIVASYDGGLYVLDGETLKLAKT
IALPGFNRVNPAGDNEHVVVSTDSGFRVFDATRQEFTDAEFKGSKPGHVVRHGGKTVLFTDGTGEVNVFDPADLSD
GKKPDGRTYTSAKPHHGVAIELAGGELVTTLGTEEKRTGALVLDKDNKEIARAENCPGVHGEAAAQGEVAGFGCED
GVLLYKDGKFTKVDAPGDYARTGNQAGSDASPILLGDYKTDPDAELERPTRISLIDTRTAKMKLVDLGTSYSFRSL

ARGPHGEALVLGTNGTLHVIDPETGKVEKKIDAVGDWTEPLDWQQPRPTLFVRDHTAYVSEPGKRQLHSIDLESGK
KLASVTLPKGTNELSGTVAGH*

SEQ ID NO: 70
Nucleotides 31332-32537 of SEQ ID NO: 1

SEQ ID NO: 71
VSWMNDVLTAVSDMNPVTRFALASVFAFAESGLGAGMAVPGEVAVLALSAGTEGTRPLLALFLVVTLSSSAGDHIG
YFLGIRYGQRMRETRLVRRIGQHHWDRAQELCHRYGARAVFLTRLLPVVRTLTPATAGVGSVRYLRFLPASLAGAA
MWSALYVSAGTLVSTSLREAESVLSTILWALLGVAAAFTLAIVWWRRRHRRRSS*

SEQ ID NO: 72
Nucleotides 32816-33427 of SEQ ID NO: 1

SEQ ID NO: 73
MELCALHSRDRDATVKTCAAGRPKRKPSYGFLGRPTAAEELAAVTSCGGGACAATTRSRA*

SEQ ID NO: 74
Nucleotides 32590-32868 of SEQ ID NO: 1

SEQ ID NO: 75
MGGSAIRTRQLTKHFGAVQALVGVDLEVPAGSVLGLLGHNGAGKTTLIQILSTVLPPSGGSAEVAGFDIVRDARRV
RACIGVTGQFAALDEHLSGLANLVLISRLLGARPREARRRAAELVEQFGLTEAADRPMRTYSGGMRRRIDLAASLV
ARPSVLFLDEPTTGLDPVSRTALWETVEGLVAEGTTVLLTTQYLDEADRLADRIAVLSSGHVVTVGTAAELKAAGT
RSVRLTFGSAADLESAEGALRLEGLGLTTDPVSRTVSLPLAATAELAGIFRILGAAGVELAELALKEPTLDDVYLS
LAESWETTSGGTVRC*

SEQ ID NO: 76
Nucleotides 34195-35154 of SEQ ID NO: 1

SEQ ID NO: 77
LTTRRTGPGTSPVADGPGWRGGGAGIGTQFRVLTGRQFRIIYGDRRIALFSLLQPIIMLMLFSQVLGRMANPEIFP
PGVRYLDYLVPALLLTTGIGSAQGGGLGLVRDMESGMMVRLRVMPVRLPLVLVARSLADLARVALQLVALLACAMG
PLGYRPAGGVSGIVGATLLALLVAWSLIWVFLALAAWLRSIEVLSSIGFLVTFPLMFASSAFVPLDILPGWLRVIA
TVNPLTYAVEASRDLALDHSALGAALAAVGTSLALLAVTGLLAVRGLRRPPGAGGPHRTP*

SEQ ID NO: 78
Nucleotides 35148-36017 of SEQ ID NO: 1

SEQ ID NO: 79
MANPFENNDGSYLVLVNDEGQYSLWPAFADVPAGWTVTFGESSRQECLDHINENWTDMRPKSLIRQMENDRTTAA*

SEQ ID NO: 80
Nucleotides 85272-85499 of SEQ ID NO: 1

SEQ ID NO: 81
MTVHDYHVTVKEQHPALFELLDPARLVAVTDEPWVTEGNEFDDDHAGRGVSYRCAQQHGEARRTGIETILGMFAGP
GGLRDMGRVLDVLGGEGLLSRVWRQLAGAGDGDSVPLVTGDLSGHMVAAALRSGLPAVRQPADRMLQRDHCLDGVL
FAYGTHHVDRSVRPRMLTEASRVLAPGGRVVLHDFAEGSPEERWFREVVHPRSLAGHAYDHFTAHEMTGYLADAGF
TDITVGPVYDPMTLTGETDESALARLVSYMTSMFGILPDGDRSNERTEAALRDIFRFSAGDLPEDVPRDEAVLELT
VRPHGNAFRAELPRIALVAHGRKP*

SEQ ID NO: 82
Nucleotides 86436-87422 of SEQ ID NO: 1

SEQ ID NO: 83
MTAQDTRTTGSDGGGRGATYHESPTYGELLRLEDLLNVAHLRDAAAPVLFLATHQSAEIWFGIVLRHLEEIRAALT
DDDPDTALHLLPRLPEIFELLVRHFDMLATLSTEEFGKIRAGLGTASGFQSAQYREIEFLCGLRXHRHISTPGFTE
TERRDCGNGPASPPWRRLRRLPDPMRQREGRERIGEALLRFDERVTVWRARHAALAERFLGPLEGTAGTAGADYLW
RVTRHRLFPPEAWGAG*

SEQ ID NO: 84
Nucleotides 87419-88153 of SEQ ID NO: 1

SEQ ID NO: 85
MDREAEAPLRAAPHATPAERAALGKAARREAPRSGHAEFSPSPRRPDPLTVLEAQSADRVPELVPIRYARMTESPF
RFYRGAAALMAADLAGTPVSGIRAQLCGDAHLLNFRLLASPERNLLFDINDFDETLPGPWEWDVKRLAASLVIAGR
ANSFTLRERAGVVRATVRSYREAMARFAGMRNLDVWYARTDAERLRTVATEQLGGRGRRNVDRALGKARSRDSLQA
FGKLAEVVDGRLRIAADPPMVVPLTDLTPGVDRDAVFRQFGSMLAGYARSLPSDRRSLLEDFALVDVARKVVGVGS
VGTRCWIVLLLGRDGGD

SEQ ID NO: 86
Nucleotides 965-1 of SEQ ID NO: 103

SEQ ID NO: 87
MIHIRAVSPPDLTDEVVGLLSADPCVLNLIVQRDAARRPDGDAIACDVLTGAANDVLHRLRAAHLDRRGSLVIEPV
DMAFSGAATEGGQRELGPLSRAPVWEQVEARIRSGGRYPPSFYLYLVIAGLIGSVGIVTNSQILIVGAMVVGPEYG
AIVSVALGIDRRHRSMVRSGLAALGVGLLLTIVVTFLFALLIRGFGLESEAFDRGLRPVSHLINTPNFFSVAVATL
AGIVGIVSLTEARTSALLGVFISVTTIPAAADIAVSTAYTSWSDVRGSAIQLVVNILVLIVVGAFALKAQRAIWQR
VRLRRDRERRIAEQA*

SEQ ID NO: 88
Nucleotides 989-1948 of SEQ ID NO: 103

SEQ ID NO: 89
VTRPGWDHEGVDTPDTPDAFPEPLPGADEAVREERATDDGTPEGRRLVRCRLCGRPLTGADSRRAGLGPSCDAKLH
PAPPDIRTRRHEVDQDPLPGT*

SEQ ID NO: 90
Nucleotides 2099-2392 of SEQ ID NO: 103

SEQ ID NO: 91
MTNPAERLVDLLDLERIEVNIFRGRSPEESLQRVFGGQVAGQALVAAGRTTDGERPVHSLHAYFLRPGRPGVPIVY
QVERVRDGRSFTTRRVTAVQEGRTIFNLTASFHRPEEAGFEHQLPPARIVPDPEELPTVAEEVREHLGALPEALER
MARRQPFDIRYVDRLRWTKDEIQDADPRSAVWMRAVGPLGDDPLVHTCALTYASDMTLLDAVRIPVEPLWGPRGYD
LASLDHAMWFHRPFRADEWFLYDQESPIATGGRGLARGRIYDRSGQLLVSVVQEGLFRRLEQ*

SEQ ID NO: 92
Nucleotides 3277-2405 of SEQ ID NO: 103

SEQ ID NO: 93
VIFVPSAGSLIRAEDRQDGGVTLIDQLPQTADPDALFEAFSSWTESQGITMYPAQEEALIEVVSGANVILSTPTGS
GKSLVAAGAHFTALAQDKVTFYTAPIKALVSEKFFDLCKLFGTENVGMLTGDASVNADAPVICCTAEVLASIALRD
GKYADIGQVVMDEFHFYAEPDRGWAWQIPLLELPQAQFVLMSATLGDVSMFEKDLTRRTGRPTSVVRSATRPVPLS
YEYRFTPITETLTELLDTRQSPVYIVHFTQAAAVERAQSLMSINMCTKEEKEKIADLIGSFRFTTKFGQNLSRYVR
HGIGVHHAGMLPKYRRLVEKLAQAGLLKVICGTDTLGVGVNVPIRTVLFTALTKYDGNRVRTLRAREFHQIAGRAG
RAGFDTAGFVVAQAPEHVIENEKALKKAGDDPKKKRKVVRKKAPEGFVAWSESTFDKLIQSEPEPLTSRFRVTHTM
LLAVIARPGNAFEAMRHLLEDNHEPRRAQLRHIRRAIAIYRSLLDGGVVEQLDTPDAEGRIVRLTVDLQQDFALNQ
PLSTFALAAFDLLDAESPSYALDMVSVVESTLDDPRQILPAQQNKARGEPVGQMKADGVEYEERMERLQEVTYPKP
LSELLWHAYDVYRTSHPWVNDHPVSPKSVIRDMYERAMTFTEFTSHYELARTEGIVLRYLASAYKALEHTIPDDVK
SEDLQDLISWLGEMVRQVDSSLLDEWEQLANPEVETAEQAQEKADEVKPVTANARAFRVLVRNAMFRRVELAALDR

AGALGELDGESGWDEDAWGEALDAYWDAHEEIGTGPDARGPKLLKIEEDPAHGLWRVWQAFADPAGDHDWGIKAEV
DLAASDEEGRAVVRVTEVGQL*

SEQ ID NO: 94
Nucleotides 5885-3312 of SEQ ID NO: 103

SEQ ID NO: 95
MMGPAHSLSGAAAWLGVGAAAAAAGHTMPWPVLVVGALICAGAALAPDLDHKSATISRAFGPVSKALCEIVDKLSY
AVYKATKSAGDPRRTGGHRTLTHTWLWAVLIGGGCSVAAITGGRWAVLVILFVHLVLAVEGLLWRAARVSSDVLVW
LLGATSAWILAGVLDKPGYGADWLFDAPGQEYMWLGLPIVLGALVHDIGDALTVSGCPILWPIPIGRKRWYPIGPP
KAMRFRAGSWVEMKVLMPAFMVLGGVGGAAALNYI*

SEQ ID NO: 96
Nucleotides 5963-6754 of SEQ ID NO: 103

SEQ ID NO: 97
MLLAELAQVSLEVAATSARSKKVALLAGLFRDAGPEDVPVVIPYLAGRLPQGRIGVGWRSLGDPVEPAAEPTLTVT
GVDARLTALAAVSGPGSQARRKEHLRALFAAATEDEQRFLRALLTGEVRQGALDALAADALARAADAPPADVRRAV
MLAGSLQEVAGVLLADGPEALAAFRLTVGRPVQPMLAHTAASVGEALDKLGACAVEEKLDGIRVQVHRDGDRIRAY
TRTLDDITDRLPELTAXVAALPAGRFI

SEQ ID NO: 98
Nucleotides 6850-8403 of SEQ ID NO: 103

SEQ ID NO: 99
VNHPVNGAGERRTTQAREGTQTVAPPRILVVGAGFAGVECVRRLERRLAPGEAQITLVTPFSYQLYLPLLPQVASG
VLTPQSVAVSLRRSRRHRTRIVPGGAIGVDTQAKVCVIRKITDEIVNEPYDYLVLAAGSVTRTFDIPGLLDNARGM
KTLAEAAYVRDHVIAQLDLADASHDEAERASRLQFVVVGGGYAGTETAACLQRLTTNAVKHYPRLDPRLIKWHLID
IAPKLMPELGDKLGQAALEVLRKRNIEVSLGVSIAEAGPEEVTFTDGRVLPCRTLIWTAGVAASPLVATLGAETVR
GRLAVTPQMRLPGADGVFSLGDAAAVPDLAKGDGAVCPPTAQHAMRQGRVLADNLIASLRHEPLKDYVHKDLGLVV
DLGGTDAVSKPLGIELRGLPAQAVARGYHWSALRTNVAKTRVMTNWLLNAVAGDDFVRTGFQSRKPATLRDFEYTD
VYLTPEQIKEHTAATVIKH*

SEQ ID NO: 100
Nucleotides 9860-8433 of SEQ ID NO: 103

SEQ ID NO: 101
VTGRDLTWTDTTSTVDRGRFPDAVTPWEDPAWRAEALAWVTEGLAAHGLTETGPRAVRLRPWSVLVRLAVAGPAPV
WFKAVPPAAAFEAGLTEALARWVPARVLAPLAVEAERGWILVPDGGPVLSEVLDGRPGAPDPGYWEEPLRQYAAMQ
RELTPYAEAIEALGVPAARPRDLPALFDRLVAGNAALPREDRVALEVLRPRVADWCEELASSGVADSLDHADLHEK
QLFAPVSGRYAFFDWGDALVGHPFCSLLVPARAARERCGPEVLPRLRDAYLEPWTGGGVTAAGLRRAVSLAWRLAA
LGRAASWGRMFPVPPGGPGVAGDAEGAHWLRELAAAPPL*

SEQ ID NO: 102
Nucleotides 10784-9921 of SEQ ID NO: 103

SEQ ID NO: 103
```
  1 GGATCCCCGC CGTCCCGGCC GAGCAGCAGG ACGATCCAGC ACCGGGTGCC GACACTGCCC
 61 ACACCGACGA CCTTCCGGGC CACGTCCACC AGCGCGAAGT CCTCCAGCAG ACTGCGCCGA
121 TCGGATGGCA GGCTGCGTGC GTACCCCGCC AGCATGGAGC CGAACTGCCG GAACACCGCG
181 TCCCGGTCCA CCCCCGGCGT CAGATCGGTC AGCGGGACGA CCATCGGCGG ATCCGCCGCG
241 ATCCGCAGCC GCCCGTCGAC CACCTCGGCG AGCTTCCCGA ACGCCTGAAG GCTGTCCCGG
301 GACCGGGCCT TCCCCAACGC CCGGTCGACA TTCCTGCGCC CCCGCCCGCC CAACTGTTCC
361 GTGGCCACCG TGCGCAGCCG CTCGGCATCC GTCCGCGCGT ACCAGACGTC CAGATTGCGC
421 ATGCCCGCGA ACCGGGCCAT CGCCTCCCGG TACGAGCGGA CCGTGGCCCG GACGACCCCG
481 GCCCGCTCCC GGAGCGTGAA GCTGTTCGCC CGCCCCGCGA TGACGAGGCT CGCCGCGAGC
541 CGCTTGACGT CCCACTCCCA GGGACCCGGC AGCGTCTCGT CGAAGTCGTT GATGTCGAAC
```

```
 601 AGGAGATTCC GCTCGGGGGA GGCCAGCAGC CGGAAGTTCA GCAGATGGGC GTCACCGCAC
 661 AACTGCGCCC TGATTCCCGA CACCGGGGTG CCGGCCAGGT CGGCGGCCAT CAGCGCGGCC
 721 GCTCCCCGGT AGAAGCGGAA CGGGGACTCC GTCATCCGGG CATAGCGGAT CGGGACCAGC
 781 TCGGGAACCC GGTCCGCCGA CTGGGCTTCG AGGACGGTCA GCGGATCGGG GCGGCGCGGC
 841 GACGGGGAGA ACTCCGCATG GCCCGACCGG GGCGCCTCAC GGCGGGCCGC CTTGCCCAGT
 901 GCCGCCCGTT CGGCCGGTGT CGCGTGCGGT GCGGCGCGCA GCGGCGCCTC GGCTTCCCGG
 961 TCCATGACGT GGCTCCTTCC GGTCTTCCTC AGGCCTGTTC GGCGATCCGG CGCTCACGGT
1021 CGCGGCGGAG GCGGACCCGC TGCCAGATCG CCCGCTGGGC CTTGAGCGCG AACGCGCCCA
1081 CCACGATCAG CACGAGGATG TTGACGACGA GCTGTATGGC CGAGCCCCGT ACGTCGGACC
1141 AGCTGGTGTA CGCCGTGGAG ACGGCGATGT CCGCGGCGGC CGGGATCGTC GTCACGGAGA
1201 TGAACACCCC GAGCAGAGCA CTGGTTCTGG CCTCGGTGAG CGACACGATC CCGACGATTC
1261 CGGCCAGGGT GGCGACGGCG ACGGAGAAGA AGTTCGGCGT GTTGATGAGA TGGGAGACGG
1321 GCCGCAGCCC CCGGTCGAAC GCCTCCGACT CCAGCCCGAA ACCCCGGATG AGGAGGGCGA
1381 AGAGGAAGGT GACCACGATG GTCAGGAGAA GGCCGACGCC CAGGGCGGCC AGCCCGCTGC
1441 GCACCATGGA CCGGTGGCGC CGGTCGATCC CCAGCGCCAC GCTGACGATG GCGCCGTACT
1501 CCGGGCCGAC GACCATCGCC CCGACGATCA GGATCTGCGA GTTGGTGACG ATGCCGACCG
1561 ACCCGATCAG ACCGGCGATG ACCAGGTAGA GGTAGAAGCT CGGCGGATAC CGGCCCCCGG
1621 ACCTGATGCG GGCCTCGACC TGTTCCCAGA CCGGCGCCCG GCTCAGCGGC CCCAGCTCGC
1681 GCTGCCCGCC CTCGGTGGCC GCGCCGGAGA AGGCCATGTC GACGGGTTCG ATGACGAGGG
1741 AGCCCCGCCG GTCGAGGTGG GCGGCGCGCA GCCGGTGCAG TACGTCGTTG GCCGCCCCCG
1801 TCAGTACGTC GCAGGCGATG GCGTCGCCGT CGGGGCGGCG CGCGGCGTCG CGCTGGACGA
1861 TCAGATTGAG CACGCACGGG TCGGCCGAGA GCAGGCCGAC GACCTCGTCG GTCAGGTCCG
1921 GCGGGCTCAC CGCGCGGATG TGGATCATGT CCATCCCGGC ACCTCCGCGG CTCCCTGCCC
1981 CGTCACACGG AGCTGTGCCC GGCAGGCGGC CCGGGGCTCA CTCCAGTAAC GCGGCACCGG
2041 CAACGTTCGG CAAACCGGCG GTCGCCCGCA CGGGCCGGGG CACCGGGGCC GCGGGCGGGT
2101 GACCCGCCCG GGCTGGGATC ATGAAGGGGT GGACACCCCC GACACACCCG ATGCCTTCCC
2161 CGAACCGCTG CCCGGGGCCG ACGAAGCGGT CCGGGAGGAG AGGGCCACCG ACGACGGGAC
2221 GCCGGAGGGC CGCCGCCTCG TCCGCTGCCG TCTCTGCGGC CGGCCCCTGA CCGGGGCCGA
2281 CTCGCGGCGG GCCGGCCTCG GCCCGTCCTG CGACGCCAAG CTGCACCCGG CGCCGCCGGA
2341 CATCCGCACC CGCCGCCACG AGGTCGACCA GGACCCGCTG CCGGGCACCT GAGCCGGAAC
2401 GGGGCTACTG CTCCAGCCGC CGGAACAGCC CCTCCTGCAC CACCGACACC AGCAGCTGCC
2461 CCGAACGGTC GTAGATCCGC CCCCGCGCCA GGCCCCGCCC GCCCGTGGCG ATCGGCGACT
2521 CCTGGTCGTA CAGGAACCAC TCGTCCGCCC GGAACGGCCG GTGGAACCAC ATGGCGTGGT
2581 CCAGGGACGC AAGGTCATAT CCGCGCGGGC CCCACAGCGG CTCCACCGGG ATACGGACCG
2641 CGTCCAGCAG CGTCATGTCG CTCGCGTACG TCAGCGCGCA CGTGTGCACC AGCGGGTCGT
2701 CGCCCAGCGG GCCCACCGCC CGCATCCACA CCGCGCTGCG CGGATCGGCG TCCTGGATCT
2761 CGTCCTTCGT CCAGCGCAGC CGGTCGACGT AACGGATGTC GAAGGGCTGG CGGCGGGCCA
2821 TCCGCTCCAG CGCCTCCGGC AGCGCGCCCA GATGCTCGCG CACCTCCTCG GCGACCGTCG
2881 GCAGCTCCTC CGGGTCCGGG ACGATCCGGG CGGGCGGCAG CTGGTGTTCG AAGCCCGCCT
2941 CCTCGGGGCG GTGGAAGGAC GCCGTCAGGT TGAAGATCGT CCGGCCCTCC TGGACCGCCG
3001 TCACCCGACG GGTGGTGAAG GACCGGCCGT CCCGCACCCG CTCCACCTGG TAGACGATCG
3061 GCACACCGGG ACGCCCCGGC CGCAGGAAAT AGGCGTGCAG CGAGTGCACC GGCCGCTCCC
3121 CGTCGGTGGT CCGGCCCGCC GCCACCAGCG CCTGGCCCGC GACCTGCCCG CCGAAGACCC
3181 GTTGCAGGGA CTCCTCCGGG CTGCGCCCCC GGAAGATGTT GACCTCGATC CGCTCCAGGT
3241 CGAGCAGGTC GACCAGACGC TCGGCCGGAT TCGTCATGCC GCACCTCTCC CGTCACACGT
3301 CAGGGTCCGC TTCACAGCTG GCCGACCTCG GTGACCCGGA CGACCGCCCG GCCCTCCTCG
3361 TCGGACGCCG CGAGGTCCAC CTCGGCCTTG ATGCCCCAGT CGTGATCGCC CGCCGGATCG
3421 GCGAACGCCT GCCAGACCCG CCACAGCCCG TGCGCCGGGT CCTCCTCGAT CTTCAGCAGC
3481 TTCGGGCCCC GCGCGTCCGG ACCGGTCCCG ATCTCCTCGT GCGCGTCCCA GTACGCGTCC
3541 AGCGCCTCGC CCCACGCGTC CTCGTCCCAC CCGGACTCGC CGTCCAGCTC GCCCAGCGCG
3601 CCGGCCCGGT CCAGCGCGGC CAGCTCCACC CGGCGGAACA TCGCGTTGCG CACCAGCACC
3661 CGGAAGGCGC GCGCGTTCGC CGTGACCGGC TTGACCTCGT CCGCCTTCTC CTGAGCCTGC
3721 TCCGCGGTCT CCACCTCGGG GTTGGCCAGC TGCTCCCACT CGTCCAGCAG ACTGGAGTCC
3781 ACCTGACGCA CCATCTCGCC CAGCCAGGAG ATCAGGTCCT GGAGGTCCTC CGACTTCACG
3841 TCGTCCGGGA TCGTGTGCTC CAGCGCCTTG TACGCGCTCG CCAGATACCG CAGCACGATG
3901 CCCTCGGTCC GGGCCAGCTC GTAGTGCGAA GTGAACTCCG TGAACGTCAT GGCCCGCTCG
3961 TACATGTCCC GGATCACCGA CTTCGGCGAC ACCGGATGGT CGTTCACCCA CGGGTGGCTC
4021 GTGCGGTACA CGTCGTACGC GTGCCACAGC AGCTCGCTCA GCGGCTTGGG GTACGTGACC
4081 TCCTGGAGCC GCTCCATCCG CTCCTCGTAC TCGACCCCGT CCGCCTTCAT CTGCCCCACG
4141 GGCTCGCCGC GCGCCTTGTT CTGCTGGGCG GGCAGGATCT GCCGGGGATC GTCCAGCGTC
4201 GACTCGACGA CCGAGACCAT GTCCAGCGCA TACGACGGCG ATTCGGCGTC CAGCAGGTCG
4261 AACGCGGCCA GCGCGAACGT GGACAGCGGC TGGTTCAGCG CGAAGTCCTG CTGGAGGTCG
4321 ACCGTCAGCC GCACGATCCG GCCCTCGGCG TCCGGGGTGT CCAACTGCTC CACCACCCCG
4381 CCGTCCAGCA GCGAGCGGTA GATGGCGATG GCCCGCCGGA TGTGCCGCAG CTGCGCCCGG
4441 CGCGGCTCGT GGTTGTCCTC CAGCAGATGC CGCATCGCCT CGAAGGCGTT GCCCGGGCGG
```

```
4501 GCGATGACCG CGAGCAGCAT CGTGTGGGTG ACCCGGAAAC GGGAGGTCAG CGGCTCCGGC
4561 TCGGACTGGA TCAGCTTGTC GAACGTCGAC TCCGACCAGG CGACGAAGCC CTCCGGGGCC
4621 TTCTTGCGGA CCACCTTGCG CTTCTTCTTC GGGTCGTCGC CCGCCTTCTT CAGCGCCTTC
4681 TCGTTCTCGA TGACATGCTC GGGGGCCTGT GCCACGACGA ACCCGGCCGT GTCGAACCCG
4741 GCCCGCCCGG CCCGGCCCGC GATCTGGTGG AACTCCCGCG CGCGCAGCGT CCGCACCCGG
4801 TTCCCGTCGT ACTTGGTGAG CGCCGTGAAC AGCACCGTAC GGATGGGGAC GTTGACGCCG
4861 ACGCCGAGCG TGTCCGTCCC GCAGATCACC TTCAGCAGCC CCGCCTGGGC CAGCTTCTCC
4921 ACCAGGCGGC GGTACTTCGG CAGCATCCCC GCGTGGTGCA CCCCGATGCC GTGGCGTACG
4981 TAACGGGAGA GGTTCTGGCC GAACTTGGTG GTGAAGCGGA AGCTGCCGAT CAGATCGGCG
5041 ATCTTCTCCT TCTCCTCCTT CGTGCACATG TTGATGCTCA TCAGCGACTG CGCCCGCTCC
5101 ACGGCCGCCG CCTGCGTGAA GTGCACGATG TAGACCGGCG ACTGCCGGGT GTCCAGCAGC
5161 TCGGTGAGCG TCTCGGTGAT CGGCGTGAAG CGGTACTCGT AGCTCAGCGG CACCGGGCGG
5221 GTCGCCGAGC GCACCACCGA GGTCGGGCGG CCGGTACGGC GGGTCAGGTC CTTCTCGAAC
5281 ATCGAGACGT CGCCGAGCGT CGCCGACATC AGCACGAACT GCGCCTGCGG CAGCTCCAGC
5341 AGCGGAATCT GCCAGGCCCA GCCCCGGTCC GGCTCGGCGT AGAAGTGGAA CTCGTCCATC
5401 ACGACCTGGC CGATGTCGGC GTACTTGCCG TCGCGCAGCG CGATGGAGGC CAGCACCTCG
5461 GCCGTACAGC AGATCACCGG GGCGTCCGCG TTGACCGAGG CGTCGCCGGT GAGCATGCCG
5521 ACGTTCTCGG TGCCGAAGAG CTTGCACAGG TCGAAGAACT TCTCCGACAC CAGCGCCTTG
5581 ATCGGAGCCG TGTAGAAGGT GACCTTGTCC TGGGCCAGCG CCGTGAAGTG CGCGCCCGCC
5641 GCCACCAGGC TCTTGCCCGA GCCGGTCGGG GTGGACAGGA TCACGTTCGC CCCGGAGACC
5701 ACCTCGATCA GCGCCTCCTC CTGAGCCGGG TACATCGTGA TGCCCTGGCT CTCGGTCCAT
5761 GAGGAGAAGG CCTCGAAGAG GGCGTCCGGG TCGGCGGTCT GGGGAAGCTG GTCGATGAGG
5821 GTCACGCCCC CATCTTGCCT GTCTTCCGCC CGGATGAGGG AACCGGCGGA CGGCACGAAG
5881 ATCACGGACG GTACGCTGCG GACTCAACCT GCCCGCGCCG CACCGGTGAT GGGCGCACGA
5941 ACCACTGGGG GCGGGACAGA CCATGATGGG ACCGGCACAC TCTCTGTCAG GGGCAGCGGC
6001 CTGGCTGGGG GTGGGCGCGG CGGCCGCCGC CGCGGGCCAC ACGATGCCCT GGCCCGTCCT
6061 CGTCGTCGGG GCGCTGATCT GCGCGGGAGC CGCACTCGCC CCCGACCTCG ACCACAAGTC
6121 CGCGACCATC TCGCGCGCCT TCGGCCCGGT CTCCAAAGCC CTCTGCGAGA TCGTCGACAA
6181 GCTCTCCTAC GCCGTCTACA AGGCCACCAA GAGCGCCGGG GACCCCCGCA GGACCGGCGG
6241 GCACCGCACC CTCACCCACA CCTGGCTGTG GGCCGTCCTC ATCGGCGGCG GCTGCTCCGT
6301 GGCGGCGATC ACCGGCGGCC GCTGGGCCGT CCTCGTGATC CTCTTCGTCC ACCTCGTGCT
6361 CGCCGTCGAG GGCCTGCTGT GGCGGGCCGC CCGCGTCTCC AGCGACGTTC TGGTGTGGCT
6421 GCTCGGCGCG ACCAGCGCGT GGATCCTGGC CGGCGTCCTG GACAAGCCCG GCTACGGGGC
6481 CGACTGGCTC TTCGACGCCC CCGGCCAGGA GTACATGTGG CTCGGCCTGC CCATCGTGCT
6541 CGGCGCCCTC GTCCACGACA TCGGCGACGC CCTCACGGTC TCGGGCTGCC CGATCCTGTG
6601 GCCCATCCCG ATCGGCCGCA AGCGCTGGTA CCCGATCGGC CCGCCGAAGG CCATGCGCTT
6661 CCGGGCCGGC AGCTGGGTGG AGATGAAGGT GCTGATGCCC GCCTTCATGG TCCTCGGGGG
6721 AGTGGGCGGG GCCGCCGCCC TCAACTACAT ATGACGCACC GCCGGTCGCG CCCGGTGCGT
6781 CTCCGGCGGG CGGCGCGTCC GGTGTCCTTC CGGCGGGCGG CGCGCTCCGC CCCGTAGCAC
6841 CATGGGCGCA TGCTGCTCGC CGAGCTCGCC CAGGTGTCCC TGGAGGTCGC CGCCACCTCC
6901 GCCCGGTCCA AGAAGGTGGC GCTCCTCGCC GGACTCTTCC GGGACGCCGG ACCCGAGGAC
6961 GTCCCCGTCG TCATCCCGTA CCTCGCCGGA CGGCTGCCCC AGGGCCGGAT CGGCGTGGGG
7021 TGGCGCTCCC TCGGCGACCC GGTGGAGCCC GCGGCGGAAC CCACCCTCAC CGTCACCGGC
7081 GTCGACGCCC GGCTGACCGC CCTCGCCGCC GTCTCGGGCC CCGGCTCCCA GGCCCGGCGC
7141 AAGGAGCACC TGCGCGCCCT CTTCGCCGCC GCCACCGAGG ACGAACAGCG CTTCCTGCGG
7201 GCCCTGCTCA CCGGCGAGGT ACGCCAGGGG GCCCTGGACG CCCTCGCCGC CGACGCCCTG
7261 GCCCGCGCCG CCGACGCCCC GCCCGCCGAC GTCCGGCGCG CCGTGATGCT CGCCGGATCG
7321 CTCCAGGAAG TCGCCGGGGT CCTCCTCGCG GACGGGCCCG AGGCGCTCGC CGCCTTCCGG
7381 CTCACCGTCG GACGGCCCGT CCAGCCGATG CTGGCGCACA CCGCCGCCTC GGTCGGCGAG
7441 GCCCTCGACA AACTGGGCGC GTGCGCGGTC GAGGAGAAGC TCGACGGCAT TCGGGTGCAG
7501 GTCCACCGCG ACGGCGACCG GATCCGCGCC TACACCCGGA CCCTCGACGA CATCACCGAC
7561 CGGCTGCCCG AGCTCACCGC CGMCGTCGCC GCCCTCCCGG CCGGCCGCTT CATCCTGGAC
7621 GGCGAGGTGA TCGCCCTGGG GGAGGACGGC AGGCCCCGGC CCTTCCAGGA GACCGCCTCC
7681 CGGGTGGGCT CGCGGCGGGA CGTGGCGGAG GCGGCGGCGC ACGTGCCCGT CGCCCCGGTC
7741 TTCTTCGACG CGCTCCTCGT CGACGACGAG GACCTGCTCG ACCTGCCCTT CACCGACCGC
7801 CACGCCGCCC TGGCCCGGCT CCTCCCCGAG CACCTGCGCG TCCGCCGCAC CCTCGTTCCC
7861 GACGCGGAGG ACCCGAAAGC CCGCGCGGCG GCCGACGCGT TCCTCACCGA CACCCTGGAA
7921 CGCGGCCACG AGGGAGTCGT CGTCAAGGAC CTCGCCGCCG CCTACAGCGC GGGCCGCCGG
7981 GGCGCGTCCT GGCTGAAGGT GAAGCCCGTG CACACCCTGG ACCTGGTGGT GCTGGCCGTC
8041 GAGTGGGGCA GCGGCCGGCG CACCGGCAAG CTCTCCAACC TGCACCTGGG CGCCCGCCGC
8101 CCCGACGGTA CGTTCGCGAT GCTCGGCAAG ACCTTCAAGG GGCTCACGGA CGCCCTGCTC
8161 GACTGGCAGA CCCAGCGCCT GGGCGAGCTG GCCACCGACG ACGACGGGCA CGTCGTCACC
8221 GTACGCCCGG AACTCGTCGT GGAGATCGCC TACGACGGAC TCCAGCGCTC CACCCGCTAC
8281 CCCGCCGGGG TCACCCTCCG CTTCGCCCGC GTCCTGCGCT ACCGCGACGA CAAGACCGCC
8341 CAGGAGGCGG ACACCGTGGA GACGGTCCTG TTCCCGGCGG CGGTGAGCGC GCCCCCGTCC
```

```
 8401 TGAAGGGGCG CGCTCGTACA GGGCCCGGCG GCTCAGTGCT TGATGACCGT CGCCGCCGTG
 8461 TGCTCCTTGA TCTGCTCGGG CGTCAGGTAG ACGTCCGTGT ACTCGAAGTC CCGCAGCGTC
 8521 GCCGGCTTGC GGGACTGGAA CCCGGTCCGT ACGAAGTCGT CACCGGCGAC CGCGTTCAGC
 8581 AGCCAGTTCG TCATGACGCG GGTCTTCGCC ACGTTCGTCC GCAGCGCCGA CCAGTGGTAG
 8641 CCCCGGGCCA CCGCCTGCGC GGGCAGCCCG CGCAGCTCGA TGCCCAGCGG CTTGGACACG
 8701 GCGTCCGTGC CGCCGAGGTC CACGACGAGC CCCAGATCCT TGTGCACGTA GTCCTTGAGC
 8761 GGCTCGTGGC GCAGCGAGGC GATCAGGTTG TCCGCCAGCA CCCGGCCCTG ACGCATCGCG
 8821 TGCTGTGCGG TGGGCGGGCA GACCGCCCCG TCGCCCTTCG CCAGATCGGG CACGGCGGCC
 8881 GCGTCGCCGA GCGAGAACAC CCCGTCCGCG CCCGGCAGTC TCATCTGCGG GGTCACGGCG
 8941 AGCCGGCCGC GTACCGTCTC CGCGCCGAGC GTGGCGACCA GCGGACTCGC GGCCACGCCG
 9001 GCGGTCCAGA TCAGCGTCCG GCAGGGCAGC ACCCGGCCGT CGGTGAACGT GACCTCCTCC
 9061 GGCCCCGCCT CGGCGATCGA CACCCCGAGC GACACCTCGA TGTTCCGCTT GCGGAGCACC
 9121 TCCAGCGCGG CCTGCCCGAG CTTGTCGCCG AGCTCCGGCA TCAGCTTCGG CGCGATGTCG
 9181 ATCAGATGCC ATTTGATCAG GCGCGGGTCA AGACGCGGAT AGTGCTTCAC CGCGTTGGTG
 9241 GTCAGACGCT GGAGACAGGC GGCCGTCTCC GTGCCCGCGT ACCCGCCGCC GACCACCACG
 9301 AACTGGAGCC GGGAGGCCCG CTCGGCCTCG TCGTGACTGG CGTCCGCCAG GTCCAGCTGG
 9361 GCGATGACGT GATCCCGTAC GTACGCGGCC TCGGCCAGCG TCTTCATCCC CCGCGCGTTG
 9421 TCCAGCAGCC CCGGGATGTC GAAGGTGCGG GTGACGCTGC CCGCCGCCAG CACGAGGTAG
 9481 TCGTACGGCT CGTTCACGAT CTCGTCCGTG ATCTTCCGGA TCACACAGAC CTTCGCCTGC
 9541 GTGTCCACGC CGATCGCCCC GCCCGGCACG ATCCTGGTCC GGTGACGGCG GCTGCGGCGC
 9601 AGCGACACCG CCACGGACTG CGGCGTGAGC ACCCCGGAGG CCACCTGGGG GAGCAGCGGC
 9661 AGATACAGCT GGTAGGAGAA CGGTGTGACG AGCGTGATCT GGGCCTCGCC CGGAGCGAGC
 9721 CTGCGCTCCA GACGGCGTAC GCACTCGACG CCTGCGAAGC CGGCGCCGAC GACGAGAATC
 9781 CTGGGTGGTG CCACGGTCTG CGTCCCTTCT CGGGCTTGCG TGGTTCTGCG CTCGCCTGCC
 9841 CCGTTTACCG GGTGATTCAC CCCTCATCCT CACCGGAGGC TCCGGCATCC GCCTCCTGGC
 9901 AGGGGTGAAA CGGGGCCCGG TCACAGGGGC GGGGCGGCCG CCAGCTCCCG CAGCCAGTGG
 9961 GCACCCTCGG CGTCCCCGGC CACGCCCGGA CCACCCGGCG GTACGGGGAA CATCCGCCCC
10021 CACGAGGCGG CCCGGCCCAG CGCCGCCAGC CGCCACGCCA GGCTCACCGC ACGGCGCAGC
10081 CCGGCCGCCG TGACGCCCCC GCCGGTCCAC GGCTCCAGAT AGGCGTCCCG CAGCCGGGGC
10141 AGCACCTCGG GACCACAGCG CTCACGGGCC GCACGGGCGG GTACCAGCAG GCTGCAGAAC
10201 GGATGGCCGA CGAGGGCGTC CCCCCAGTCG AAGAAGGCGT ACCGCCCGGA CACGGGCGCG
10261 AACAGCTGCT TCTCGTGCAG ATCGGCGTGG TCCAGCGAGT CCGCCACCCC CGACGACGCC
10321 AGCTCCTCGC ACCAGTCGGC CACCCGGGGC CGCAGCACCT CCAGCGCCAC CCGGTCCTCC
10381 CGGGGCAGCG CGGCGTTCCC CGCGACCAGC CGGTCGAACA GCGCGGGAAG GTCGCGCGGC
10441 CGGGCCGCCG GAACCCCCAG GGCCTCGATC GCCTCCGCGT ACGGGGTCAG CTCCCGCTGC
10501 ATCGCGGCGT ACTGGCGCAG CGGCTCCTCC CAGTAGCCGG GGTCAGGGGC GCCGGGACGC
10561 CCGTCGAGGA CCTCCGACAG CACCGGGCCG CCGTCCGGGA CGAGTATCCA GCCGCGTTCC
10621 GCCTCGACGG CGAGCGGGGC CAGCACCCGG GCCGGGACCC AGCGCGCCAG CGCCTCGGTG
10681 AGCCCCGCCT CGAAGGCCGC GGCGGGCGGA ACGGCCTTGA ACCAGACGGG CGCGGGCCCG
10741 GCGACGGCCA GCCGCACCAG CACCGACCAG GGACGCAGCC GCACCGCCCG GGGGCCCGTC
10801 TCCGTCAGCC CGTGAGCGGC GAGGCCCTCG GTCACCCAGG CGAGGGCCTC CGCCCGCCAG
10861 GCCGGGTCCT CCCAGGGCGT CACGGCGTCC GGGAAGCGCC CCCGGTCCAC GGTCGAGGTC
10921 GTGTCGGTCC AGGTCAGGTC TCTTCCGGTC ACGGCGGTCG TGGTCGTGCC GGGGCCGTCA
10981 CGGCGGTCGT GGTCACGGCA TCCGGGGCCG CGTCGGGCAT GGGCATCTCG TCTCCGCGCA
11041 TCCGATCATG GGATCACCGG CCCCGGCGCG TGCGCACCGC AATTTCCGGG AACACCCGTT
11101 CCCGTCCTGC CCGGATCGGC TGTCCTCCCC CCTCCGGCCC CTGGAACGGC GGGAGTTCGG
11161 CCGCCCGCCC CGTGCGAGGA TGCTGTGGTG ACCACCTCGC CCTCCTCGCC CGTGGCCGAC
11221 GACTCTTCCG TGTCTTCCGT GGACGACGCC CCGCCCCGCG ACCAGGGGCT GAGCTCCCGG
11281 GCCGCGGCGG TACTCGTCTT CGGGTCCTCC GCCGCGGTCC TCGTGGTCGA GATCGTCGCC
11341 CTGCGGCTGC TCGCCCCGTA CCTCGGCCTC ACCCTGGAGA CCAGCACGCT GGTGATCGGC
11401 ATCGCGCTGA CCGCCATCGC CCTGGGTTCC TGGCTGGGCG GGCGCATCGC GGACCAGGTC
11461 GATCCGCACC GGCTCATCGC CCCCGCGCTC GGGGTGTCGG GCGTGGGCGT CGCGCTCACC
11521 CCGCTCCTGC TCCGTACCAC CGCGGAGTGG TCTCCCGCGC TGCTCCTGCT GGTCGCTTCG
11581 GCGACCCTCC TGGTGCCGGG CGCGCTGCTC TCCGCGGTGA CCCCGTTCGT GACGAAGTTG
11641 CGGCTCACCA GCCTCGCCGA GACCGGGACG GTCGTCGGGC GGCTGTCGGG CGTCGGCACC
11701 TTCGGAGCCA TCGTCGGCAC GGTGCTCACC GGATTCGTCC TGGTCACGCG GCTGCCCGTC
11761 AGCTCCATCC TGATCGGCCT CGGCACGCTG CTGGTGCTCG GGGCGGCCCT CGTCGGATGG
11821 CAGGCCCGGC GGTGGCGGCG CGCCACGGCC GTGGCCCTCG CCACCGTCGT CGCGGGCACT
11881 CTCGCCACCG GGTTCGCTCC CGGCGGCTGC GACGCGGAGA CCCGCTACCA CTGCGCCCGG
11941 GTTCGTCGCG GACCCCGACC GGGGACAGCG GGCCGCACCC CTCGTCCTCG GCACGGGCCT
12001 GCGCCACTCC TACGTCGATG TCGAGGACCC CGAGTACCTG AAGTTCGCGT ACGTACGCGC
12061 CTTCGCCTCC GTGGTCGACA CGGCCTTCCC CGAGGGCGAG CCGCTGACCG CCCACCACAT
12121 CGGGGGCGGC GGCCTCACCT TCCCCCGCTA CCTCGCGGCC ACCCGCCCCG GAACCCGCAG
12181 CCTCGTCTCC GAGATCGACC CCGGGGTCGT CCGCATCGAC CGCGACCGGC TCGGCCTCGG
12241 CACCCCTGCC GCGACCGGCA TCGACGTACG CGTCGAGGAC GGGCGTCTCG GCCTGCGGCG
```

```
12301 GCTGGACGCG GGCAGCCACG ACCTGGTCGT CGGCGACGCC TTCGGAGGCG TCAGCGCGCC
12361 CTGGCACCTC ACGACGTCCC AGGCACTCAA GGACGTACGC CGGGCGCTCG ACGCGGACGG
12421 CCTGTACGTC ACCAACCTCA TCGACCACGG CCGGCTCGCC TTCGCCCGCG CCGAGGTCGC
12481 CACCCTCGCC GCGACCTTCC CGCATGTCGC GCTGCTCGGG CAGCCCGCGG ACATCGGCCT
12541 GGACCCCACG GCTTCGAGCA TCGGCGGCAA CATGGTGGTC GTCGCCTCCG CCCGGCCGGT
12601 CGACGCCCCC GCCATCCAGA AAGCCATGGA CGCCCGGGAC GTCGGCTGGA GGATCGCCAC
12661 CGGCGACACC CTCACCACCT GGACGGGGAA CGCCCGGGTG CTCACCGACG ACCACGCGCC
12721 CGTCGACCAA CTCCTCCAGC CCCACCCCGT CCCATCGGCC CGGTAAGGCC CGAACGGGCC
12781 CGATGATCCC GCCCGAACGC CCCGGTAACG CACGAACGGC CCGGTGATCC CCGSCCGTTC
12841 GCGCGGGGAT CACCGGGCCG TTCGGCCAAG ACGCCTCACC CGTGCCAGGA CCGCCACAGC
12901 GACGCGTACG CGCCGCCCGC CGCCACCAGC TCGTCATGGC TGCCCAGTTC ACTGATCCGG
12961 CCGTCCTCCA CGACCGCGAT CACATCCGCG TCGTGCGCGG TGTGCAGCCG GTGCGCGATC
```

CLAIMS

We claim:

1. An isolated nucleic acid molecule comprising a nucleic acid sequence that encodes a thioesterase or thioesterase domain, wherein a gene encoding the thioesterase or thioesterase domain is derived from a bacterial daptomycin biosynthetic gene cluster.

2. The nucleic acid molecule according to claim 1, wherein the bacterial daptomycin biosynthetic gene cluster is derived from *Streptomyces*.

3. The nucleic acid molecule according to claim 2, wherein the bacterial daptomycin biosynthetic gene cluster is derived from *S. roseosporus*.

4. The nucleic acid molecule according to claim 3, wherein the molecule is an allelic variant of a nucleic acid sequence comprising SEQ ID NO: 3, the thioesterase-encoding domain of SEQ ID NO: 3, or SEQ ID NO: 6.

5. The nucleic acid molecule according to claim 1, comprising a nucleic acid sequence which encodes the amino acid sequence GXSXG, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids.

6. The nucleic acid molecule according to claim 5, wherein the nucleic acid sequence encodes an amino acid sequence comprising the amino acid sequence GWSFG or GTSLG.

7. An isolated nucleic acid molecule comprising a nucleic acid sequence that encodes a thioesterase or a thioesterase domain, wherein the nucleic acid sequence is selected from the group consisting of:

(a) a nucleic acid sequence of *dptD*;

(b) a nucleic acid sequence of *dptH*;

(c) a nucleic acid sequence encoding the amino acid sequence of SEQ ID NO: 7;

(d) a nucleic acid sequence encoding the amino acid sequence of SEQ ID NO: 8;

(e) a nucleic acid sequence comprising the nucleic acid sequence of SEQ ID NO: 3;

(f) a nucleic acid sequence comprising the nucleic acid sequence of SEQ ID NO: 6;

(g) a nucleic acid sequence encoding a thioesterase domain of DptD, wherein said nucleic acid sequence comprises at least a portion of a nucleic acid molecule selected from *dptD*, SEQ ID NO: 3 or a nucleic acid molecule encoding SEQ ID NO: 7;

(h) a nucleic acid sequence encoding an amino acid sequence comprising the amino acid sequence GWSFG or GTSLG;

(i) a nucleic acid sequence comprising the nucleic acid sequence selected from the group consisting of

(1) nucleotides 78488-78511 of SEQ ID NO: 1,

(2) nucleotides 79898-79930 of SEQ ID NO: 1,

(3) nucleotides 80453-80488 of SEQ ID NO: 1,

(4) nucleotides 80558-80581 of SEQ ID NO: 1,

(5) nucleotides 80654-80677 of SEQ ID NO: 1,

(6) nucleotides 81050-81064 of SEQ ID NO: 1,

(7) nucleotides 81623-81646 of SEQ ID NO: 1,

(8) nucleotides 83117-83149 of SEQ ID NO: 1,

(9) nucleotides 83669-83704 of SEQ ID NO: 1,

(10) nucleotides 83774-83797 of SEQ ID NO: 1,

(11) nucleotides 83870-83893 of SEQ ID NO: 1,

(12) nucleotides 84257-84271 of SEQ ID NO: 1,

(13) nucleotides 80033-80320 of SEQ ID NO: 1, and

(14) nucleotides 83255-83542 of SEQ ID NO: 1;

(j) a nucleic acid sequence encoding an amino acid sequence selected from the group consisting of

(1) amino acids 144-151 of SEQ ID NO: 7,

(2) amino acids 614-624 of SEQ ID NO: 7,

(3) amino acids 799-810 of SEQ ID NO: 7,

(4) amino acids 834-841 of SEQ ID NO: 7,

(5) amino acids 866-873 of SEQ ID NO: 7,

(6) amino acids 998-1002 of SEQ ID NO: 7,

(7) amino acids 1189-1196 of SEQ ID NO: 7,

(8) amino acids 1687-1697 of SEQ ID NO: 7,

(9) amino acids 1871-1882 of SEQ ID NO: 7,

(10) amino acids 1906-1913 of SEQ ID NO: 7,

(11) amino acids 1938-1945 of SEQ ID NO: 7,

(12) amino acids 2067-2071 of SEQ ID NO: 7,

(13) amino acids 659-754 of SEQ ID NO: 7, and

(14) amino acids 1733-1828 of SEQ ID NO: 7;

(k) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05;

(l) a nucleic acid sequence encoding an amino acid sequence D-L-X-X-G-X_{1-33}-K-X_{1-22}-T-X-G-X_{1-23}-V-X_{1-7}-I, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(m) a nucleic acid sequence encoding an amino acid sequence D-A-X-X-W-X_{1-37}-T-X_{1-20}-T-X-T-X_{1-21}-G-X_{1-7}-V, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(n) a nucleic acid sequence comprising at least 50% sequence identity to the nucleic acid sequence of any one of (a) to (k); and

(o) a nucleic acid sequence, wherein a nucleic acid molecule comprising said sequence selectively hybridizes to the complementary strand of a nucleic acid molecule comprising the nucleic acid sequence of any one of (a) to (k).

8. The nucleic acid molecule according to claim 7, wherein the homologous molecule exhibits at least 60% sequence identity to the nucleic acid sequence of any one of (a) to (k).

9. The nucleic acid molecule according to claim 8, wherein the sequence identity is at least 70%.

10. The nucleic acid molecule according to claim 9, wherein the sequence identity is at least 80%.

11. The nucleic acid molecule according to claim 10, wherein the sequence identity is at least 90%.

12. The nucleic acid molecule according to claim 11, wherein the sequence identity is at least 95%.

13. An isolated nucleic acid molecule comprising a part of a nucleic acid sequence that encodes a thioesterase, wherein said part is at least 13 nucleotides, and wherein the nucleic acid sequence is derived from a gene from a bacterial daptomycin biosynthetic gene cluster.

14. The nucleic acid molecule according to claim 13, wherein the nucleic acid sequence is selected from the group consisting of:

(a) a nucleic acid sequence encoding DptD;

(b) a nucleic acid sequence encoding DptH;

(c) a nucleic acid sequence encoding an amino acid sequence of SEQ ID NO: 7;

(d) a nucleic acid sequence encoding an amino acid sequence of SEQ ID NO: 8;

(e) a nucleic acid sequence comprising SEQ ID NO: 3;

(f) a nucleic acid sequence comprising SEQ ID NO: 6;

(g) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05;

(h) a nucleic acid sequence encoding an amino acid sequence GXSXG, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(i) a nucleic acid sequence comprising the nucleic acid sequence selected from the group consisting of

(1) nucleotides 78488-78511 of SEQ ID NO: 1,

(2) nucleotides 79898-79930 of SEQ ID NO: 1,

(3) nucleotides 80453-80488 of SEQ ID NO: 1,

(4) nucleotides 80558-80581 of SEQ ID NO: 1,

(5) nucleotides 80654-80677 of SEQ ID NO: 1,

(6) nucleotides 81050-81064 of SEQ ID NO: 1,

(7) nucleotides 81623-81646 of SEQ ID NO: 1,

(8) nucleotides 83117-83149 of SEQ ID NO: 1,

(9) nucleotides 83669-83704 of SEQ ID NO: 1,

(10) nucleotides 83774-83797 of SEQ ID NO: 1,

(11) nucleotides 83870-83893 of SEQ ID NO: 1,

(12) nucleotides 84257-84271 of SEQ ID NO: 1,

(13) nucleotides 80033-80320 of SEQ ID NO: 1, and

(14) nucleotides 83255-83542 of SEQ ID NO: 1;

(j) a nucleic acid sequence encoding an amino acid sequence selected from the group consisting of

(1) amino acids 144-151 of SEQ ID NO: 7,

(2) amino acids 614-624 of SEQ ID NO: 7,

(3) amino acids 799-810 of SEQ ID NO: 7,

(4) amino acids 834-841 of SEQ ID NO: 7,

(5) amino acids 866-873 of SEQ ID NO: 7,

(6) amino acids 998-1002 of SEQ ID NO: 7,

(7) amino acids 1189-1196 of SEQ ID NO: 7,

(8) amino acids 1687-1697 of SEQ ID NO: 7,

(9) amino acids 1871-1882 of SEQ ID NO: 7,

(10) amino acids 1906-1913 of SEQ ID NO: 7,

(11) amino acids 1938-1945 of SEQ ID NO: 7,

(12) amino acids 2067-2071 of SEQ ID NO: 7,

(13) amino acids 659-754 of SEQ ID NO: 7, and

(14) amino acids 1733-1828 of SEQ ID NO: 7;

(k) a nucleic acid sequence encoding an amino acid sequence D-L-X-X-G-X_{1-33}-K-X_{1-22}-T-X-G-X_{1-23}-V-X_{1-7}-I, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(l) a nucleic acid sequence encoding an amino acid sequence D-A-X-X-W-X_{1-37}-T-X_{1-20}-T-X-T-X_{1-21}-G-X_{1-7}-V, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(m) a nucleic acid sequence comprising at least 70% sequence identity to a nucleic acid sequence of any one of (a) to (j); and

(n) a nucleic acid sequence, wherein a nucleic acid molecule comprising said sequence selectively hybridizes to the complementary strand of a nucleic acid molecule comprising the nucleic acid sequence of any one of (a) to (j).

15. The nucleic acid molecule according to claim 14, wherein the part comprises at least 14 nucleotides of the nucleic acid sequence.

16. The nucleic acid molecule according to claim 15, wherein the part comprises at least 17 nucleotides of the nucleic acid sequence.

17. The nucleic acid molecule according to claim 16, wherein the part comprises at least 20 nucleotides of the nucleic acid sequence.

18. The nucleic acid molecule according to claim 17, wherein the part comprises at least 25 nucleotides of the nucleic acid sequence.

19. The nucleic acid molecule according to either of claims 13 or 14, wherein the part encodes an amino acid sequence comprising the amino acid sequence GWSFG or GTSLG.

20. The nucleic acid molecule according to any one of claims 11-19, wherein the part encodes a polypeptide with thioesterase activity.

21. The nucleic acid molecule according to any one of claims 11-19 that is an oligonucleotide from 14 to 60 nucleotides in length.

22. An isolated nucleic acid molecule comprising a nucleic acid sequence encoding a daptomycin non-ribosomal peptide synthetase (NRPS) or subunit thereof from *Streptomyces*, wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

23. The nucleic acid molecule according to claim 22, wherein the daptomycin NRPS or subunit thereof is from *S. roseosporus.*

24. An isolated nucleic acid molecule comprising a nucleic acid sequence encoding a daptomycin non-ribosomal peptide synthetase (NRPS) or subunit thereof from *Streptomyces roseosporus*, wherein the nucleic acid molecule encodes a polypeptide selected from the group consisting of DptA, DptB, DptC and DptD, wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160,

pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

25. The nucleic acid molecule according to claim 24, wherein the nucleic acid molecule encodes a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7.

26. The nucleic acid molecule according to claim 23, wherein the nucleic acid molecule is selected from the group consisting of *dptA*, *dptB*, *dptC* and *dptD* or wherein the nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3.

27. The nucleic acid molecule according to claim 24, wherein the nucleic acid molecule comprises a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05.

28. An isolated nucleic acid molecule that encodes a daptomycin NRPS or subunit thereof, wherein the isolated nucleic acid molecule selectively hybridizes to a reference nucleic acid molecule that encodes a daptomycin NRPS or subunit thereof, wherein the reference nucleic acid molecule comprises a nucleic acid sequence selected from the group consisting of:

(a) a nucleic acid sequence selected from the group consisting of *dptA*, *dptB*, *dptC* or *dptD*;

(b) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05; and wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

29. The isolated nucleic acid molecule according to claim 28, wherein the nucleic acid molecule hybridizes under conditions selected from the group consisting of low stringency conditions, moderate stringency conditions and high stringency conditions.

30. An isolated nucleic acid molecule that encodes a daptomycin NRPS or subunit thereof, wherein the isolated nucleic acid molecule comprises a nucleic acid sequence that has at least 50% sequence identity to a nucleic acid sequence selected from the group consisting of:

(a) a nucleic acid sequence selected from the group consisting of *dptA, dptB, dptC* or *dptD*;

(b) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05; and

wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

31. The nucleic acid molecule according to claim 30, wherein the homologous molecule exhibits at least 60% sequence identity to the nucleic acid sequence of any one of (a) to (e).

32. The nucleic acid molecule according to claim 31, wherein the sequence identity is at least 70%.

33. The nucleic acid molecule according to claim 32, wherein the sequence identity is at least 80%.

34. The nucleic acid molecule according to claim 33, wherein the sequence identity is at least 90%.

35. The nucleic acid molecule according to claim 34, wherein the sequence identity is at least 95%.

36. An isolated nucleic acid molecule that encodes a daptomycin NRPS or subunit thereof, wherein the isolated nucleic acid molecule is an allelic variant of a nucleic acid molecule that comprises a nucleic acid sequence selected from the group consisting of:

(a) a nucleic acid sequence selected from the group consisting of *dptA, dptB, dptC* or *dptD*;

(b) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05; and

wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

37. An isolated nucleic acid molecule that encodes at least one domain from a daptomycin NRPS, wherein the nucleic acid molecule comprises a part of a nucleic acid sequence of at least 14 nucleotides, selected from the group consisting of:

(a) a nucleic acid sequence selected from the group consisting of *dptA, dptB, dptC* or *dptD*;

(b) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05; and

wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

38. An isolated nucleic acid molecule that encodes at least one module from a daptomycin NRPS, wherein the nucleic acid molecule comprises a part of a nucleic acid sequence of at least 14 nucleotides selected from the group consisting of:

(a) a nucleic acid sequence selected from the group consisting of *dptA*, *dptB*, *dptC* or *dptD*;

(b) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05; and

wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

39. An isolated nucleic acid molecule comprising a part of a nucleic acid sequence, wherein said part is at least 14 nucleotides, selected from the group consisting of:

(a) a nucleic acid sequence selected from the group consisting of *dptA*, *dptB*, *dptC* or *dptD*;

(b) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05; and

wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

40. The nucleic acid molecule according to claim 39, wherein the part comprises at least 17 nucleotides of the nucleic acid sequence.

41. The nucleic acid molecule according to claim 40, wherein the part comprises at least 20 nucleotides of the nucleic acid sequence.

42. The nucleic acid molecule according to claim 41, wherein the part comprises at least 25 nucleotides of the nucleic acid sequence.

43. The nucleic acid molecule according to claim 42, wherein the part comprises at least 50 nucleotides of the nucleic acid sequence.

44. The nucleic acid molecule according to any one of claims 39-43 that is an oligonucleotide from 14 to 60 nucleotides in length.

45. A vector comprising the nucleic acid molecule according to any one of claims 1-44.

46. The vector according to claim 45, wherein the vector comprises expression control sequences controlling the transcription of the nucleic acid molecule.

47. The vector according to claim 46 wherein the expression control sequences control the expression of the nucleic acid molecule in a prokaryotic cell.

48. A host cell comprising the nucleic acid molecule according to any one of claims 1-44.

49. A host cell comprising the vector according to any one of claims 44-47.

50. A method for producing a polypeptide selected from the group consisting of a thioesterase, a daptomycin NRPS, and a daptomycin NRPS subunit, comprising the step of culturing the host cell according to claims 48 or 49 under conditions in which the polypeptide is produced, optionally comprising the step of isolating the polypeptide.

51. An isolated nucleic acid molecule comprising an expression control sequence derived from a gene encoding a thioesterase or a daptomycin NRPS derived from a bacterial daptomycin biosynthetic gene cluster, wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

52. The nucleic acid molecule according to claim 51, wherein the bacterial daptomycin biosynthetic gene cluster is derived from *Streptomyces*.

53. The nucleic acid molecule according to claim 52, wherein the bacterial daptomycin biosynthetic gene cluster is derived from *S. roseosporus*.

54. The nucleic acid molecule according to claim 53, wherein the expression control sequence is derived from the daptomycin NRPS or DptH.

55. The nucleic acid molecule according to claim 53, wherein the nucleic acid molecule comprises all or a part of the nucleic acid sequence of SEQ ID NO: 2 or SEQ ID NO: 5.

56. The nucleic acid molecule according to claim 55, wherein said part is at least 30 nucleotides in length.

57. The nucleic acid molecule according to claim 56, wherein said part is at least 50 nucleotides in length.

58. The nucleic acid molecule according to claim 57, wherein said part is at least 100 nucleotides in length.

59. The nucleic acid molecule according to claim 58, wherein said part is at least 200 nucleotides in length.

60. A vector comprising the nucleic acid molecule according to any one of claims 51-59.

61. The vector according to claim 60, wherein the nucleic acid molecule is operatively linked to a second nucleic acid molecule so as to regulate the expression of the second nucleic acid molecule.

62. The vector according to claim 61, wherein the second nucleic acid molecule encodes a polypeptide derived from a bacterial daptomycin biosynthetic gene cluster selected from the group consisting of a thioesterase, a daptomycin NRPS and a daptomycin NRPS subunit.

63. The vector according to claim 61, wherein the second nucleic acid molecule is a heterologous nucleic acid molecule.

64. An isolated polypeptide comprising an amino acid sequence that encodes a thioesterase or a fragment thereof, wherein said thioesterase is derived from a bacterial daptomycin biosynthetic gene cluster.

65. An isolated polypeptide comprising an amino acid sequence that encodes a daptomycin NRPS, a subunit thereof, a module thereof or a domain thereof, wherein said daptomycin NRPS is derived from a bacterial daptomycin biosynthetic gene cluster.

66. The polypeptide according to claim 64 or 65, wherein the bacterial daptomycin biosynthetic gene cluster is derived from *Streptomyces*.

67. The polypeptide according to claim 66, wherein the bacterial daptomycin biosynthetic gene cluster is derived from *S. roseosporus*.

68. The polypeptide according to claim 65, wherein the polypeptide is a thioesterase or fragment thereof, which comprises the amino acid sequence GXSXG, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids.

69. The polypeptide according to claim 68, wherein the thioesterase or fragment thereof comprises the amino acid sequence GWSFG or GTSLG.

70. An isolated polypeptide comprising an amino acid sequence that encodes a thioesterase or a fragment thereof, wherein the polypeptide comprises an amino acid sequence selected from the group consisting of:

(a) an amino acid sequence from a thioesterase domain of DptD;

(b) an amino acid sequence of DptH;

(c) the amino acid sequence of a thioesterase domain of SEQ ID NO: 7;

(d) the amino acid sequence of SEQ ID NO: 8;

(e) an amino acid sequence encoded by a thioesterase-encoding region of the nucleic acid sequence of SEQ ID NO: 3;

(f) an amino acid sequence encoded by a coding region of the nucleic acid sequence of SEQ ID NO: 6;

(g) the amino acid sequence GXSXG, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(h) an amino acid sequence encoded by the nucleic acid sequence selected from the group consisting of

(1) nucleotides 78488-78511 of SEQ ID NO: 1,

(2) nucleotides 79898-79930 of SEQ ID NO: 1,

(3) nucleotides 80453-80488 of SEQ ID NO: 1,

(4) nucleotides 80558-80581 of SEQ ID NO: 1,

(5) nucleotides 80654-80677 of SEQ ID NO: 1,

(6) nucleotides 81050-81064 of SEQ ID NO: 1,

(7) nucleotides 81623-81646 of SEQ ID NO: 1,

(8) nucleotides 83117-83149 of SEQ ID NO: 1,

(9) nucleotides 83669-83704 of SEQ ID NO: 1,

(10) nucleotides 83774-83797 of SEQ ID NO: 1,

(11) nucleotides 83870-83893 of SEQ ID NO: 1,

(12) nucleotides 84257-84271 of SEQ ID NO: 1,

(13) nucleotides 80033-80320 of SEQ ID NO: 1, and

(14) nucleotides 83255-83542 of SEQ ID NO: 1;

(i) an amino acid sequence selected from the group consisting of

(1) amino acids 144-151 of SEQ ID NO: 7,

(2) amino acids 614-624 of SEQ ID NO: 7,

(3) amino acids 799-810 of SEQ ID NO: 7,

(4) amino acids 834-841 of SEQ ID NO: 7,

(5) amino acids 866-873 of SEQ ID NO: 7,

(6) amino acids 998-1002 of SEQ ID NO: 7,

(7) amino acids 1189-1196 of SEQ ID NO: 7,

(8) amino acids 1687-1697 of SEQ ID NO: 7,

(9) amino acids 1871-1882 of SEQ ID NO: 7,

(10) amino acids 1906-1913 of SEQ ID NO: 7,

(11) amino acids 1938-1945 of SEQ ID NO: 7,

(12) amino acids 2067-2071 of SEQ ID NO: 7,

(13) amino acids 659-754 of SEQ ID NO: 7, and

(14) amino acids 1733-1828 of SEQ ID NO: 7;

(j) an amino acid sequence encoded by a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05;

(k) an amino acid sequence D-L-X-X-G-$X_{1\text{-}33}$-K-$X_{1\text{-}22}$-T-X-G-$X_{1\text{-}23}$-V-$X_{1\text{-}7}$-I, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(l) an amino acid sequence D-A-X-X-W-$X_{1\text{-}37}$-T-$X_{1\text{-}20}$-T-X-T-$X_{1\text{-}21}$-G-$X_{1\text{-}7}$-V, wherein each X is independently selected from any one of the twenty naturally-occurring L-amino acids;

(m) an amino acid sequence comprising at least 50% sequence identity to the amino acid sequence of any one of (a) to (j); and

(n) an amino acid sequence encoded by a nucleic acid sequence, wherein a nucleic acid molecule comprising said nucleic acid sequence selectively hybridizes to the complementary strand of a nucleic acid molecule encoding the amino acid sequence of any one of (a) to (j).

71. The polypeptide according to claim 70, wherein the polypeptide has thioesterase activity.

72. The polypeptide according to claim 71, wherein the polypeptide exhibits at least 60% identity to the amino acid sequence of any one of (a) to (j).

73. The polypeptide according to claim 72, wherein the sequence identity is at least 70%.

74. The polypeptide according to claim 73, wherein the sequence identity is at least 80%.

75. The polypeptide according to claim 74, wherein the sequence identity is at least 90%.

76. The polypeptide according to claim 75, wherein the sequence identity is at least 95%.

77. The polypeptide according to claim 70, wherein the polypeptide is a polypeptide fragment, a fusion polypeptide, a polypeptide derivative, a polypeptide analog, a mutein or a homologous polypeptide of a naturally-occurring thioesterase derived from a daptomycin biosynthetic gene cluster.

78. The polypeptide according to claim 77, wherein the polypeptide is a polypeptide fragment comprising at least 5 contiguous amino acids.

79. The polypeptide according to claim 78, wherein the fragment comprises at least 10 amino acids.

80. The polypeptide according to claim 79, wherein the fragment comprises at least 20 amino acids.

81. The polypeptide according to claim 80, wherein the fragment comprises at least 50 amino acids.

82. The polypeptide according to claim 77, which is a fusion protein comprising at least 10 amino acids from the thioesterase.

83. The polypeptide according to claim 82, comprising at least 50 amino acids from the thioesterase.

84. The polypeptide according to claim 82, wherein the fusion protein comprises the amino acid sequence encodes thioesterase activity.

85. An isolated polypeptide according to any one of claims 65-67, wherein the polypeptide has an amino acid sequence selected from the group consisting of

(a) an amino acid sequence encoded by a nucleic acid sequence selected from the group consisting of *dptA*, *dptB*, *dptC* or *dptD*;

(b) an amino acid sequence selected from the group consisting of DptA, DptB, DptC or DptD;

(c) a nucleic acid sequence encoding the amino acid sequence of a polypeptide selected from the group consisting of SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13 and SEQ ID NO: 7;

(d) a nucleic acid sequence selected from the group consisting of SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14 and SEQ ID NO: 3; and

(e) a nucleic acid sequence from an *S. roseosporus* nucleic acid sequence from BAC clone B12:03A05.

86. An isolated polypeptide that is encoded by the nucleic acid molecule according to any one of claims 28-36.

87. An isolated polypeptide that is encoded by the nucleic acid molecule according to claim 37.

88. An isolated polypeptide that is encoded by the nucleic acid molecule according to claim 38.

89. An antibody that selectively binds to the polypeptide according to any one of claims 64-88.

90. The antibody according to claim 89 that is an intact immunoglobulin; an antigen-binding portion thereof that is Fab, Fab', $F(ab')_2$, Fv, dAb or a CDR fragment; a single-chain antibody; a chimeric antibody; a diabody; or a polypeptide comprising at least a portion of the immunoglobulin sufficient to confer specific antigen binding to the polypeptide.

91. The antibody according to claim 90, wherein the antibody is a neutralizing antibody.

92. The antibody according to claim 90, wherein the antibody is an activating antibody.

93. The antibody according to claim 90, wherein the antibody is a monoclonal antibody or a polyclonal antibody.

94. A method for preparing an antibody that selectively binds to the polypeptide according to any one of claims 64-88, comprising the steps of

a) immunizing a non-human animal with the polypeptide; and

b) isolating the antibody.

95. A method for determining if a sample contained a nucleic acid molecule encoding a thioesterase, a daptomycin NRPS or a daptomycin NRPS subunit, comprising the steps of

a) providing a nucleic acid molecule according to any one of claims 1-43;

b) contacting the nucleic acid molecule with the sample under selective hybridization conditions; and

c) determining if the nucleic acid molecule selectively hybridized to a nucleic acid molecule in the sample.

96. A method for amplifying a second nucleic acid molecule encoding a thioesterase or a portion thereof from a sample comprising the second nucleic acid molecule, comprising the steps of

a) providing a first nucleic acid molecule, wherein the first nucleic acid molecule comprises the nucleic acid sequence according to any one of claims 1-12 and comprises at least 10 contiguous nucleotides of the nucleic acid sequence;

b) contacting the first nucleic acid molecule with the sample comprising the second nucleic acid molecule under conditions in which the first and second nucleic acid molecules will selectively hybridize to each other; and

c) amplifying the second nucleic acid molecule using polymerase chain reaction (PCR).

97. A method to produce daptomycin comprising the steps of

a) introducing a nucleic acid molecule comprising a daptomycin biosynthetic gene cluster or a portion thereof sufficient to direct the synthesis of daptomycin into a host cell; and

b) culturing the host cell under conditions in which daptomycin is produced.

98. The method according to claim 97, wherein the nucleic acid molecule is derived from *Streptomyces*.

99. The method according to claim 98, wherein the nucleic acid molecule is derived from *S. roseosporus*.

100. The method according to claim 99, wherein the nucleic acid molecule comprises the entire daptomycin biosynthetic gene cluster.

101. The method according to claim 97, wherein the host cell is *S. lividans*.

102. The method according to claim 101, wherein the host cell is *S. lividans* TK64.

103. The method according to claim 97, further comprising the step of isolating the daptomycin.

104. A method to increase the production of daptomycin by a cell comprising the steps of

a) providing a host cell that expresses daptomycin;

b) introducing a nucleic acid molecule into a neutral site of a chromosome of said host cell, wherein the introduction of the nucleic acid molecule results in increased production of daptomycin by a cell compared to the cell without the nucleic acid molecule; and

c) culturing the host cell under conditions in which daptomycin is produced;

wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

105. The method according to claim 104, wherein the host cell is *S. roseosporus* or *S. lividans* comprising the daptomycin biosynthetic gene cluster.

106. The method according to either of claims 104 or 105, wherein the nucleic acid molecule is selected from the group consisting of *NovA,B,C, dptA, dptB, dptC, dptD, dptE, dptF, dptG, dptH, and* fatty acyl-CoA ligase from the daptomycin biosynthetic gene cluster and any combination of two or more nucleic acid molecules thereof.

107. The method according to either of claims 104 or 105, wherein the nucleic acid molecule is a daptomycin resistance gene.

108. The method according to claim 106, further comprising the step of introducing a daptomycin resistance gene into the host cell.

109. The method according to either of claims 104 or 105, wherein the nucleic acid molecule is the entire daptomycin biosynthetic gene cluster or BAC clone B12:03A05.

110. The method according to claim 109, further comprising the step of introducing a daptomycin resistance gene into the host cell.

111. A method for producing a modified daptomycin, comprising the steps of

a) providing a cell comprising a daptomcyin biosynthetic gene cluster or a portion thereof sufficient to direct the synthesis of daptomycin into a host cell;

b) modifying or replacing one or more modules of the daptomycin biosynthetic gene cluster or portion thereof to alter the amino acid that is incorporated into the modified daptomycin; and

c) culturing the host cell under conditions in which modified daptomycin is produced.

112. The method according to claim 111, wherein one or more modules specifying incorporation of aspartate is modified to specify incorporation of asparagine or 3-methyl-glutamate.

113. The method according to claim 111, wherein the module is replaced by a module derived from a non-ribosomal peptide synthetase other than the daptomycin biosynthetic gene cluster.

114. The method according to claim 113, wherein the module specifying incorporation of L-kynurnine is replaced by a module specifying incorporation of L-tryptophan.

115. A method for producing a modified daptomycin, comprising the steps of

a) providing a cell comprising a daptomycin biosynthetic gene cluster or a portion thereof sufficient to direct the synthesis of daptomycin into a host cell;

b) inserting or deleting one or more modules of the daptomycin biosynthetic gene cluster or portion thereof to insert or delete one or more amino acids in the cyclic peptide of the modified daptomycin; and

c) culturing the host cell under conditions in which modified daptomycin is produced.

116. The method according to claim 115, further comprising the step of altering one or more adenylation domains.

117. The method according to claim 115, wherein the module is inserted directly upstream from a thioesterase module.

118. A method to create a modified daptomycin, comprising the steps of

a) providing a cell comprising a daptomcyin biosynthetic gene cluster or a portion thereof sufficient to direct the synthesis of daptomycin into a host cell;

b) inserting or translocating a thioesterase domain to the end of an internal module to delete one or more amino acids in the cyclic peptide of the modified daptomycin; and

c) culturing the host cell under conditions in which modified daptomycin is produced.

119. The method according to claim 118, wherein the thioesterase domain is translocated.

120. A method to produce a hybrid non-ribosomal peptide synthetase (NRPS) or polyketide synthetase (PKS) comprising the steps of

a) providing a nucleic acid molecule encoding a thioesterase from a daptomycin biosynthetic gene cluster; and

b) linking the nucleic acid molecule encoding the thioesterase to a nucleic acid molecule encoding a natural or synthetic NRPS or PKS.

121. The method according to claim 120, wherein the nucleic acid molecule encoding the thioesterase is linked to nucleic acid sequences from the daptomycin biosynthetic gene cluster and one or more other NRPS or PKS.

122. The method according to claim 120, wherein the nucleic acid molecule encoding the thioesterase is linked to nucleic acid sequences not derived from the daptomycin biosynthetic gene cluster.

123. The method according to claim 120, wherein the method is used to produce a novel cyclic peptide or linear peptide.

124. A method to produce a cyclic thioester comprising the steps of providing a pantetheine-peptide thioester intermediate to a thioesterase derived from a daptomycin biosynthetic gene cluster.

125. The method according to claim 124, wherein the thioesterase is derived from a nucleic acid molecule comprising SEQ ID NO: 3 or SEQ ID NO:6.

126. A method to determine whether a lipopeptide is an antibiotic, comprising the steps of

a) providing a linear thioester tethered to a cleavable resin;

b) adding a thioester to cyclize the thioester;

c) encapsulating the lipopeptide with a test strain of bacteria;

d) cleaving the cyclic thioester from the resin; and

e) determining if the cyclic thioester has antibiotic activity against the test strain.

127. The method according to claim 126, wherein the resin is a photocleavable resin and the cleaving step is performed using light.

128. The method according to claim 126, wherein the method is used in high throughput screening.

129. The method according to claim 126, wherein the peptide is attached to the resin via a lipid, alkyl or polyether linker.

130. A method for identifying a thioesterase, comprising the steps of

a) providing a linear thioester peptide tethered to a cleavable resin, wherein the thioester peptide, when cyclized, has antibiotic activity;

b) providing a DNA library in an expression vector that does not lyse a host cell;

c) introducing the DNA library into a host cell that is resistant to the cyclized peptide product;

d) encapsulating the host cell comprising the DNA library and the linear thioester peptide into a matrix to form a macrodroplet;

e) incubating the macrodroplet such that the host cell expresses the polypeptide from the DNA library;

f) placing the macrodroplet on an appropriate target lawn and cleaving the thioester peptide;

g) determining whether the thioester peptide in each macrodroplet has antibiotic activity; and

h) isolating the DNA from the macrodroplet that has antibiotic activity.

131. A method to cyclize peptides, comprising the steps of

a) providing a peptide that contains – and C-terminal amino acid residues that are recognized by a thioesterase derived from a daptomycin biosynthetic gene cluster; and

b) contacting the peptide with the thioesterase under conditions in which cyclization occurs.

132. The method according to claim 131, wherein the peptide is produced by an NRPS or a PKS.

133. The method according to claim 132, wherein the peptide is located within a cell.

134. The method according to claim 133, wherein the thioesterase is encoded by a nucleic acid molecule that has been introduced into the cell.

135. The method according to claim 134, wherein the nucleic acid molecule encoding the thioesterase is operatively linked to a heterologous promoter.

136. The method according to claim 135, wherein the nucleic acid molecule encoding the thioesterase is operatively linked to its naturally-occurring promoter.

137. A nucleic acid molecule comprising a nucleic acid sequence selected from the group consisting of SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 and 102 or encoding a polypeptide having an amino acid sequence selected from the group consisting of SEQ ID NOS: 19, 21, 23, 25, 27,

29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 and 101;

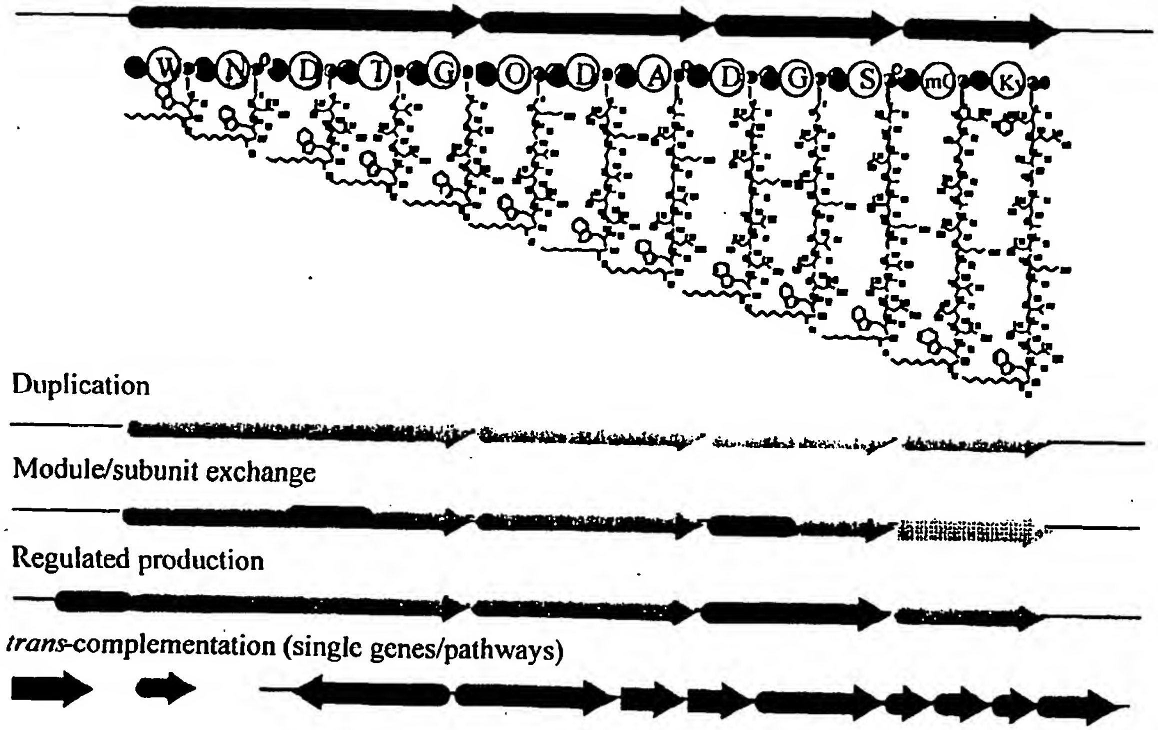
wherein said nucleic acid molecule is not pRHB153, pRHB157, pRHB159, pRHB160, pRHB166, pRHB168, pRHB172, pRHB599, pRHB602, pRHB603, pRHB680, pRHB613 or pRHB614.

138. A polypeptide comprising an amino acid sequence selected from the group consisting of SEQ ID NOS: 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99 and 101 or encoded by a nucleic acid molecule selected from the group consisting of SEQ ID NOS: 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100 and 102.

139. An antibody that binds to the polypeptide according to claim 138.

Figure 1

Manipulations of *Dpt* genes



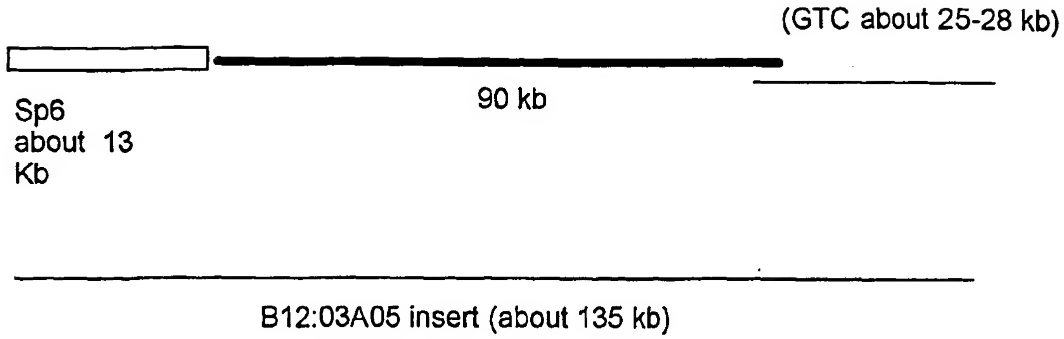
Duplication

Module/subunit exchange

Regulated production

trans-complementation (single genes/pathways)

Figure 2A


(GTC about 25-28 kb)
90 kb
Sp6
about 13
Kb
B12:03A05 insert (about 135 kb)

90 KB Fragment

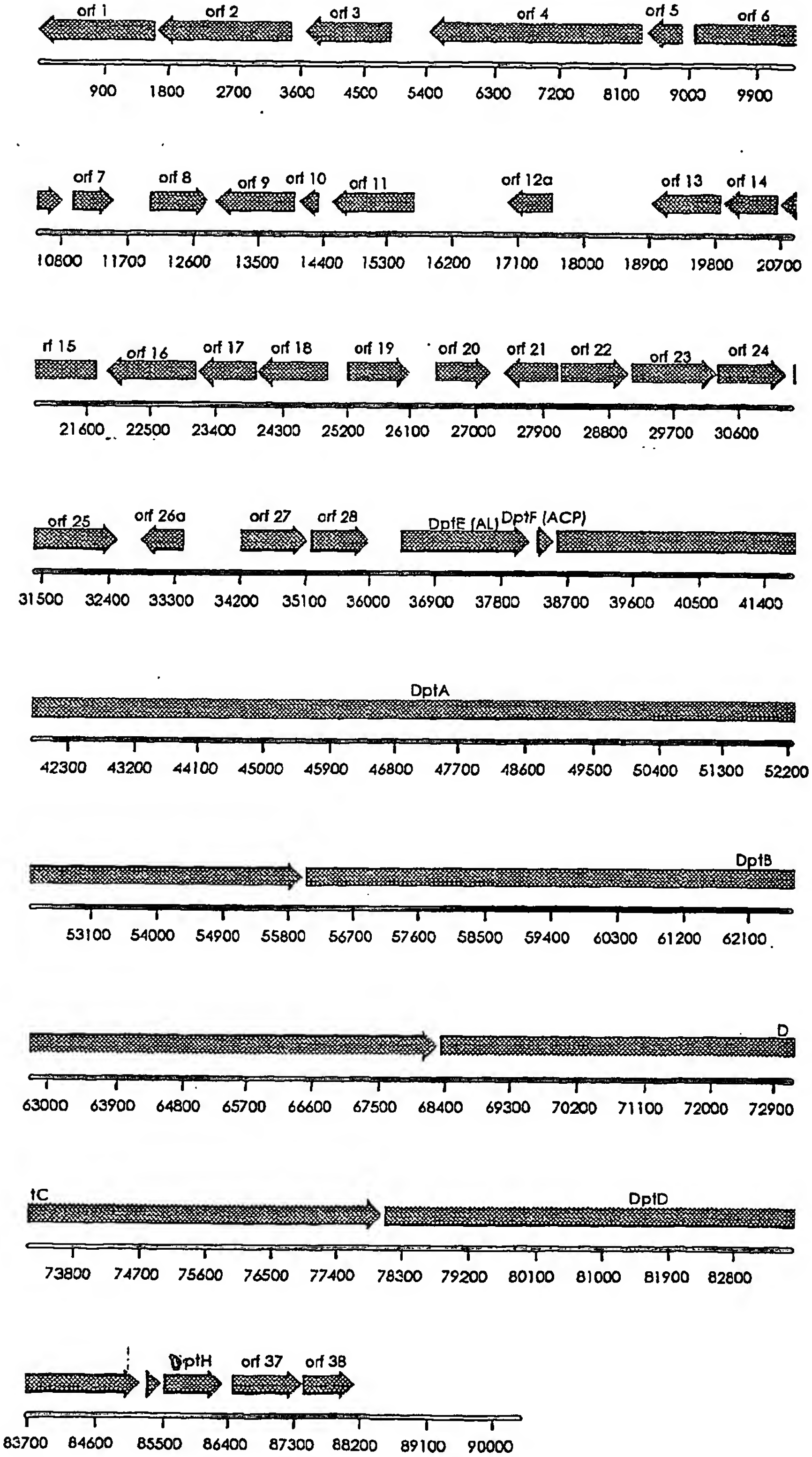


Figure 2B

SP6 Fragment

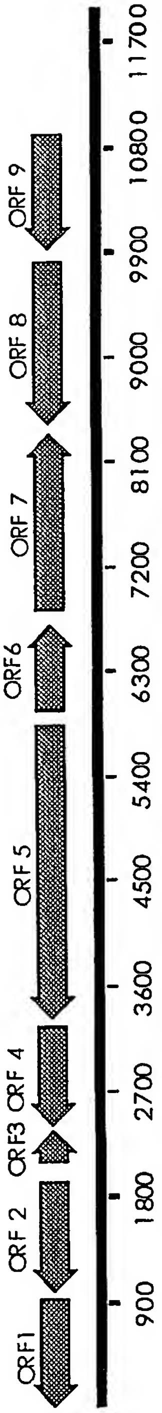
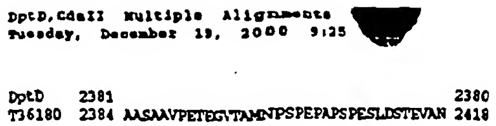


Figure 2C

```
DptD,CdaII  Multiple  Alignments
Tuesday,  December  19,  2000  9:25

ClustalW (v1.4) multiple sequence alignment

2 Sequences Aligned           Alignment Score = 7705
Gaps Inserted = 22            Conserved Identities = 1223

Pairwise Alignment Mode: Slow
Pairwise Alignment Parameters:
    Open Gap Penalty = 10.0   Extend Gap Penalty = 0.1
    Similarity Matrix: blosum

Multiple Alignment Parameters:
    Open Gap Penalty = 10.0   Extend Gap Penalty = 0.1
    Delay Divergent = 40%     Gap Distance = 8
    Similarity Matrix: blosum

Processing time: 18.0 seconds


DptD        1 MTQRAHEDILPLTPLQEGLLFHSVYDEQSVDVYTVQVVVDLEGPVDPEALRAAAAALLRRHANLRAAFRYERLQRPVQIIPREVAVPWEHTDVAKLEGAEQKAEIERLLHDQRWRRFDLT  120
T36180      1 MAKSGLEDILPLSPLQEGMLFHNLFDEEELDAYNVQVFIDLEGGTDPERLRRAGQALLERHANLRAAFRHEGLKRPVQLIPRRVVLPWGEEDLSGVAEPEREAAAERVAERDRWTRFDLS  120

DptD      121 APPLLRFLLVRTGHDRHRFALTFHHILMDGWSMPVLLRELITLYRTG-DETALPWVRPYRDYLAWISRRDRDEAGRAWSKALAGVDEATLVAPGADRAAEPPLWTESRLEPDLAATLAAR  239
T36180    121 RPPLIRFTLVRLGPARHRLLLTLHHILADGWSMPILLRELMTLYTVHGDGTALPRVRPYRDYLGHLGRDRDAARQAWTEAFAGLDAPSIVAPGRGALTAAPERIDFSEDEAASAALTRF  240

DptD      240 AREFGVTLNTLVQAAWALVLGRLTGRDDVVFGVTVSGRPPELAGVEDMVGLFINTVPLRAELLPHESLADFTVRLQREQIQLLDHQYERLAVIQRLAGRTELFDTVMVFENYPVAAASS-  358
T36180    241 ARSNGLTVNTVIQGCWGLVLSHLTGRDDVVFGVTVSGRPPELPGIDTMVGLFMNTLPLRVRLRPAETLTGFLRRLQGEQARLIDHQWVGLAEIQRWAGSGELFDTAMVFENYPLNSSRGR  360

DptD      359 -AGADGPAAEPRVADVHVRDAMHYPLGLLVLPGPPLRLRFGHRPSALPAERVTTIRDSLVRALELMADQPDLAVGRADILGEEEKQHLLTGLNDTHRDVPPLTVPGHIEAQAARTPGRPA  477
T36180    361 PPGAAPDADLPTVLGVRSKDQMHYPLGLLALPRETLRFSLGYLPQVFDPARVEAVIAAFRRALRTVLDAPDTRVGAVALLDPEVRGTVLERWSGSDDVRPAERFTDLFEEQVARTPGKTA  480

DptD      478 VHARDGELSYAELNARANRLARHLAAAGVGPEQYVTLLLPLSARMVVAALAVHKTGAAYVPVDPEYPADRIAYMLGDIGPALVLTDS---RSAAAMPAGPARVLTLDDDALDTGVRALPE  594
T36180    481 LIAPDGRLTYAELDAAANRLARRLVELGVGPERHVAVAVGRRTELVVGHLAVLKAGGAYVPVDPEYPPDRIRHMIQDADPALVLTTSDVDDRIGEECCGPLTFVMIDDPNTGTSLGRHSGT  600

DptD      595 HDLGTDGIAPLPD-QPAYVIYTSGSTGRPKGVVILHRSVTGYLLRT-IEEYPEAAGKAFVHSPVSFDLTVGALYAPLVSGGCLRLGSFTDDKILDLGE---DSPTFMKATPSHLAVLDSL  709
T36180    601 ALTDADRAAFLLPGHPAYVIYTSGTTGRPKGVVVEHRALSAFVRHCRSSQAPDISGLSVMQASASFDQSVGSLHAPLISGGCVRLTDLRALAETAGSEPGFHRATFMKGTPSHLALLATM  720

DptD      710 PDEISPTGAITLGGEQLLSETLDPWRARHPGVTVFNVYGPTETTINCAEHRIAPGTTLPPGPVPIGRPLWNTRLYVLDGGLRVVPTGVAGELYVAGAGLARGYLGRPGLTAERFVACPFG  829
T36180    721 PPEVAPSGTLTLGGEELRGEILAPWREAAGDVTVVNVYGPTEATGHCLEHWIAPDRTVEPGPVPIGTPHEGVRVYVLDSALRPVAPGLDGEVYLAGVQLARGYLGRGGLTAERFTADPPG  840

DptD      830 APGERMYRTGDLVRWRTDGTLEFVGRVDDQVKVRGFRIELGEVEATVAATPGVARAIVAVREDRPGDQRLVAYVTPADVDPTGGLPSAVTAHAAARLPAYMVPSAVVVLHEVPLTPNGKI  949
T36180    841 APGSRMYRTGDVARWNEAGELVFAGRADRQVKLRGYRIELGEIEAAVAGGPGVRQAAVVLREDRPGDQRLVAYVVP---DPGHWDEAAARARLALSLPDFMMPSAFVALDALPLSPNGKL  957

DptD      950 NRAALPAPEAVSGAGFRAPGTAREEVLCGLFAEVLGLERVGTADDFFELGGHSLLATRLVSRVRSVLGVELGVRALFDAPTPGRLDRLLGERSGAPVRAPLTARERTGRDPLSYAQQRLW 1069
T36180    958 DRAALPAPTYTGRTAGRAPRTPAEEILCDLYAEVLSLPGVTVDDDFFDLGGHSLLATRLVSRVRTTLGAELSIRQFFEAPTPAALAVVLAG--AGRARAALTARPRPERLPLSYAQQRLW 1075

DptD     1070 FLHELEGHGATYNIPLALRLTGPLDVTALEAALTDVVARHESLRTLIARDGTGTAMQHILPTGDPRARITLEAVFLHRDELAGRLAEAARHPFDLTAEIPVRATVFRTERDDHTLLVVTH 1189
T36180   1076 FLHLLEGPSPTYNIPTVLRLSGPLFPDALRAALLDVVGRHESLRTTFTEDERG-ARQVVHPADG--VRPVFETAESTEADYEADLARAARHAFDLGAEIPVRARLLRLSEREHVLLLLVH 1192

DptD     1190 HIASDRWSREPFLRDLSAAYAARRAHSAPELPPLSVQYADYAAWQRDVLGTEDDGTSEMAQQLAHWRGRLAGLPQGLDLPTDRPRRPDVGRRGGRCRLEIPAALHRDIVTLARVTSTTVF 1309
T36180   1193 HIASDAWSRGPLAQDLTAAYTARCAGDAPAWQPLPVQYADYALWQQELLGDDTDPDTLAGRQLAYWKQQLAGLPERLDLPTDRPRPATADHTGDRVEFALPADLHTRLTELARATDTTLY 1312

DptD     1310 MVVQAALAGLLSRLGAGTDIPIGTPIAGRTDEATEHLIGFFVNTLVLRTDVSGDPTFAELLARVRATDLDAYAHQDVPFERLVEVLNPERSLLRHPLFQILLAFQNTEDRSISDRPGTLL 1429
T36180   1313 MVLQAALATLLTRHGAGEDIPIGTPVAGRTDDATDHLVGFFVNTLVLRTDTSGNPTFRDLLTRVRDTDLTAYTHQDLPFERLVEALNPTRSLTHHPLFQVVLSLRSTAPRRADGEGAPAL 1432

DptD     1430 PDLQVTEQP--LDAGTAKFDLAPAFTERPPEKGEPSGITGIVEYHADLYDEGTVRQIADCFVQFLDAAVHAPGTRVDAVGLLPEHTLHQLLTRSRGTVTGLPPATLPELFEARVAAHPGH 1547
T36180   1433 PGGLRVSGTGGAAATAAKVDLGPSVTERRAADHTPDGVAGVLDFRTDLFDRGTAQGLVDRLVRVLADAAAHPDRPLSRIDVLGPRERHRVVEEWNATAKGLAPATLPELFERHVRERPGA 1552

DptD     1548 IAVEVAGRRPATTTYDALNRRANRLARLLTDRGVRPEQRVAIALPRSADLVTAMLGILKAGAVCVPVDPAYPDDRIAHMAADAAPALLIASAATRDRMLPTGIPVLDLDDP--AVTAALA 1665
T36180   1553 EAVVAG---DTSLSYAELNARANRLARLLVARGAGPERLVALALPRSAELPVAVLAVAKAGAAYLPLDPAHPAERIAGTLDDAAPVALLTTAAVAAGLPDTDVPRLLLDEEPAAGGGEDA 1669

DptD     1666 AAPDGNPRGTGLLPAHPAYVIYTSGSTGTPKGVVVTHEGIPALAATQQEALRAGPGDRVLQLVSTSFDASVWDLCSALLSGATLVLAPDADLFGDELAAALTAHRITHVTLPPAALAAVP 1785
T36180   1670 ADLTDADRLAPLLPGHPAYVIYTSGTTGRFKGVTVTHSGLPALLDIFTSQLDVVPGSRVLHHLSPAFDGGFWELAMGLLTGAALVVVEPGTVPGPALAALAVRHRVTHAAITPAVLQLIP 1789

DptD     1786 AGAAPPRLTVTVTGDVCGPQLVDRHAGGERRILNGYGPTEVTVGATYAVCERTGDGAPVPIGAPWPDQRVYVLEHRLRPVPAGCVGELYVAGAGLARGYLGRPGQTAERFVADPFGAPGE 1905
T36180   1790 EGALPAGTTLVVAAETCPPELVARWSAG-RLMRNSYGPTETTVCATMSAPLAG--AAVPPIGRPIADTAGYVLDDALQPVPPGVPGELYVRGPGLARGYLGRPSLTAGRFVACPFGPAGG 1906

DptD     1906 RMYRTGDLARRRSDGHLLFEGRADTQVKIRGFRVELAEIEAALASHPGVEDAVVTVYDDGLGDQRLVAYVTGGP-GTPSAAALRAHLASRLPRHMVPGDVLTLDALPLTANGKVDRTALP 2024
T36180   1907 VMYRTGDLVRHRADGDLEYLGRTDTQVKLRGHRVEPAEIEAVTAGLPGVAQAAVLVREDTPGDRRLVGYVVPDAGASVDPGALRQALRGSLPEYMVPAALVVLDALPLTTNGKLDHRALP 2026

DptD     2025 GPGTQTAAPGRAPQSPQERVLCALFADVLGRETVGVDEGFFDLGGHSLLATRLAARVRAALGVEISVRTLFEAPTPALLASACTADAAAYDPPETVLPLRRTGSRPPLFCVHAGHGLSWA 2144
T36180   2027 APEYRTVE-GRSPRTPREEALCRLFAEVLGLELVGLDDGFFDLGGHSLLATRLVERVRAELGEELGVRDLFAAPTVADLAVRLAARGGR-EPMERLLPLRAAGTARPVFCVHPGSGMSWC 2144

DptD     2145 YAGLLSHLDADVPVYGLQARRLTAPGGLPGSVEEMAEDYAGEIRRLCPDGPYRLLGWSFGGTVAHAVATRLQQQGHTVELLAVLDAYPVTGARPDAEVDEQRIVADYLAQLGSPVAPERL 2264
T36180   2145 YSGLVRHLPPGIPVYGLQAAGLDGDGPLPATLQEMAAEYADLVRQTQPEGPYRLLGHSLGGNVAFAMARELAARGQEVELLAFLDAYPRR-AGAGPEAPLAEVFAHNLRDAGFDVAEEEL 2263

DptD     2265 EG-DAHLPEFLEFVRRTDGPARDFDACRILAMKDVFLNNARLTRRFTPGVPTGDMVPFASARP-GSEQAAERVGLWHPHVTGDLDLHLIDCAHEEMTDP-AALTRIGPVLAARLGACTN  2380
T36180   2264 TGGRFPTARYRAFLNAAGDPHGRLDEAELAAVLEVFHNNAALMRGHTPGTYDGDVLVLAAERADGDKLARRGAESWRPHVRGRIERVGVDADHLGLVQSCAALAVIGRALAGRLDPATGH 2383
```



```
DptD,CdaII  Multiple  Alignments
Tuesday,  December  19,  2000  9:25

DptD    2381                                    2380
T36180  2384 AASAAVPETEGVTAMNPSPEPAPSPESLDSTEVAN 2418
```

```
DptB/Cda  prot  alignmen|      tiple  Alignments
Tuesday,  December  19, |      3:37  AM

ClustalW (v1.4) multiple sequence alignment

2 Sequences Aligned             Alignment Score = 955
Gaps Inserted = 1               Conserved Identities = 145

Pairwise Alignment Mode: Slow
Pairwise Alignment Parameters:
    Open Gap Penalty = 10.0   Extend Gap Penalty = 0.1
    Similarity Matrix: blosum

Multiple Alignment Parameters:
    Open Gap Penalty = 10.0   Extend Gap Penalty = 0.1
    Delay Divergent  = 40%    Gap Distance = 8
    Similarity Matrix: blosum

Processing time: 0.4 seconds

                          10        20        30        40        50        60
DptH protein      MRATSFRMIQVNGARIACSDSG------CGDPVLMIAGTGSTGRVWDAYQVPDLHAAGFRT
T36181                MPVLTVNGIRINYYDDAPPAGAQNAPAVLLVMGSGGSGRAWHLHQVPALVAAGFRV
                      .. *** **    *              **.. *.* .** *   *** * *****

                          70        80        90       100       110       120
DptH protein      ITFTNRGVPPSDECERGFTLADLAADTAALIEQVAGGPCRVVGTSLGAQVAQEVALARPD
T36181            ISFDNRGIAPSEECPGGFGIDDLVADTAALVEELRLGPCRVAGISMGAHIAQELALSRPD
                  *.* ***. **.**  **  . ** ******.*..  ***** * *.**..***.**.***

                         130       140       150       160       170       180
DptH protein      LVTQAVFMATRGRTDAMRAAATRAAAALYDSGVELPPAYAAAVRALQNLSPHTLRDRHQV
T36181            LVDRLVLMATRARPDALREALCRAEMELYDQGIRLPAAYEAVVQAMQNLSPRTLDNDVQA
                  **  .  * **** * **.* * .**   *** *. ** ** * *.*.*****.**    *
                                           .

                         190       200       210       220       230       240
DptH protein      EDWLPLFEYAERDGPGVRAQLELGLLPDRLADYRDITVPCLVIAFEDDVVTPPYLGREVA
T36181            RDWLDVLELTRRSGAGYRAQLGVRVDGDRREAYRGIRAATRVVAFQDDLIAPPHLGREVA
                   ***  . * . * * * ****  . .  **   ** *   . *.**.**...** ******

                         250       260       270
DptH protein      DAIPGARFETVPRCGHYGYLEDASAVNKILRDFFRTSN
T36181            DAIPGAEYELVPDCGHYGYLESPDAVNKSLVEFLRRN
                  ****** .* ** *********   **** * .* *
```

Figure 5A

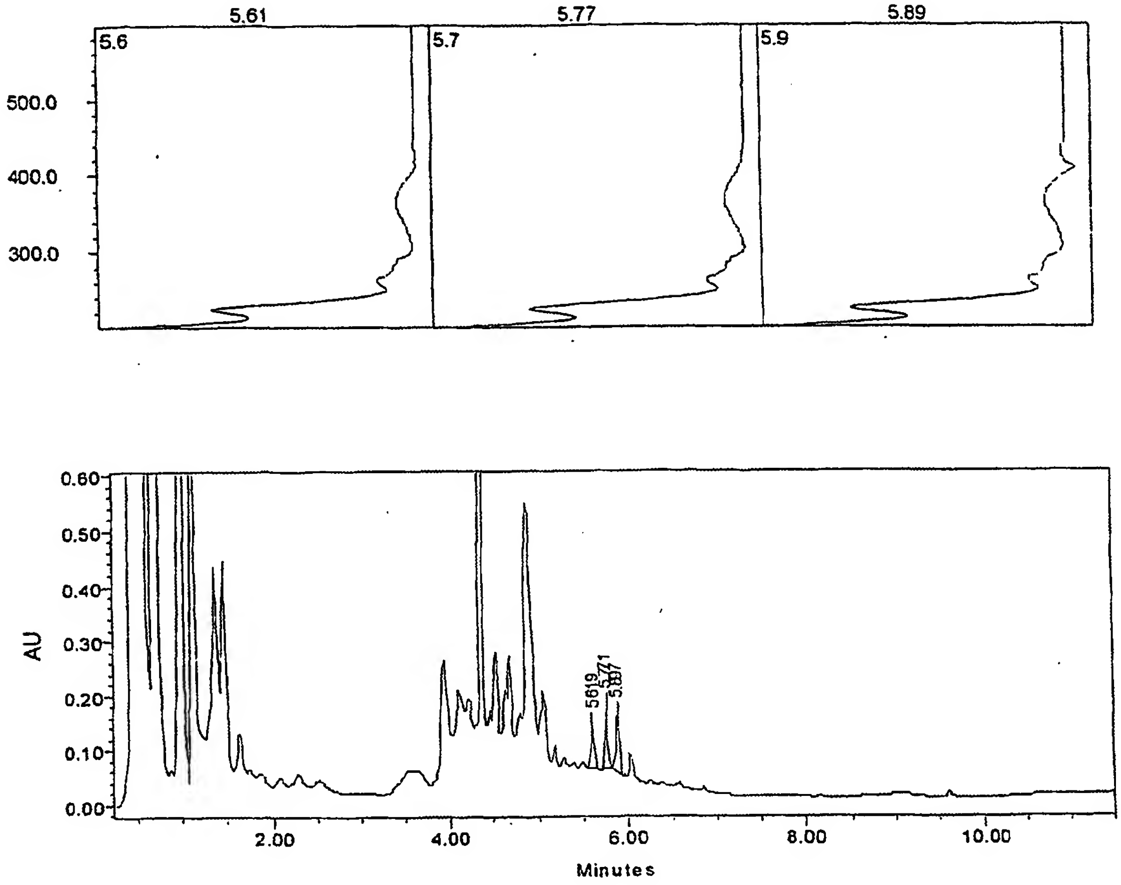

5.61
5.77
5.89
5.6
5.7
5.9
500.0
400.0
300.0
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AU
5.619
5.771
5.897
2.00
4.00
6.00
8.00
10.00
Minutes

Figure 5B

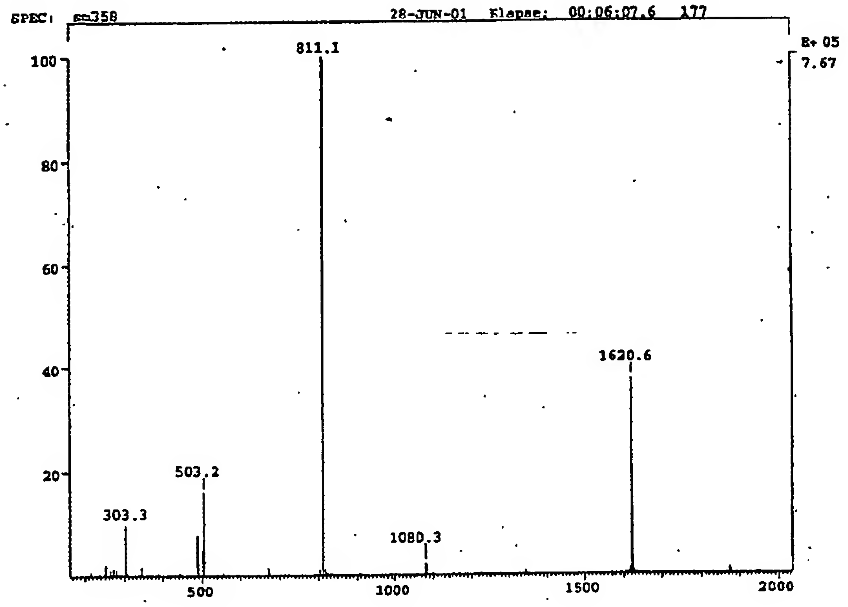
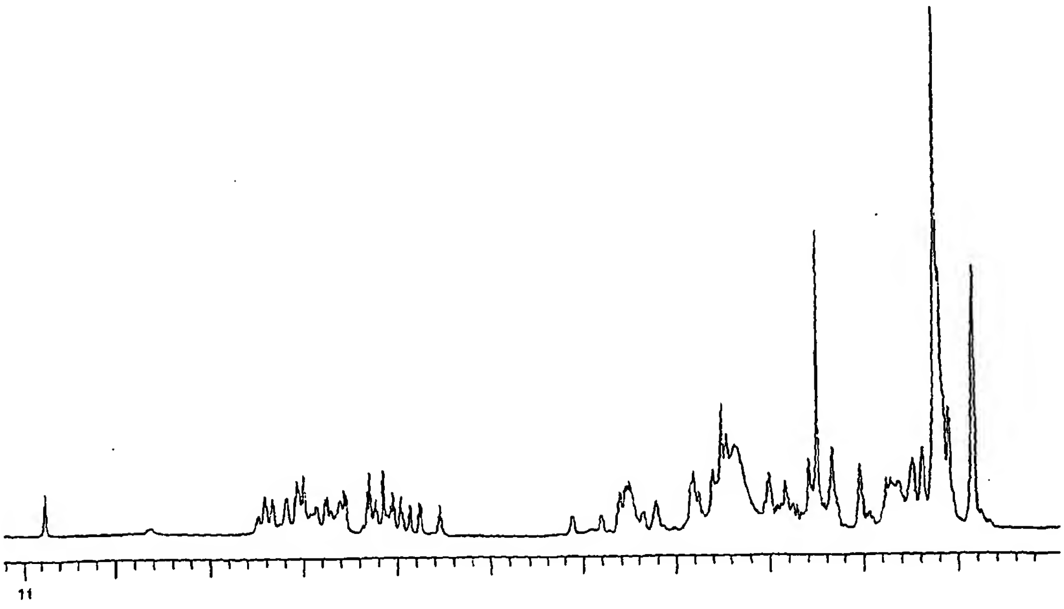

SPEC: sm358
28-JUN-01 Elapse: 00:06:07.6 177
E+ 05
7.67
100
80
60
40
20
811.1
1620.6
503.2
303.3
1080.3
500
1000
1500
2000

Figure 5C



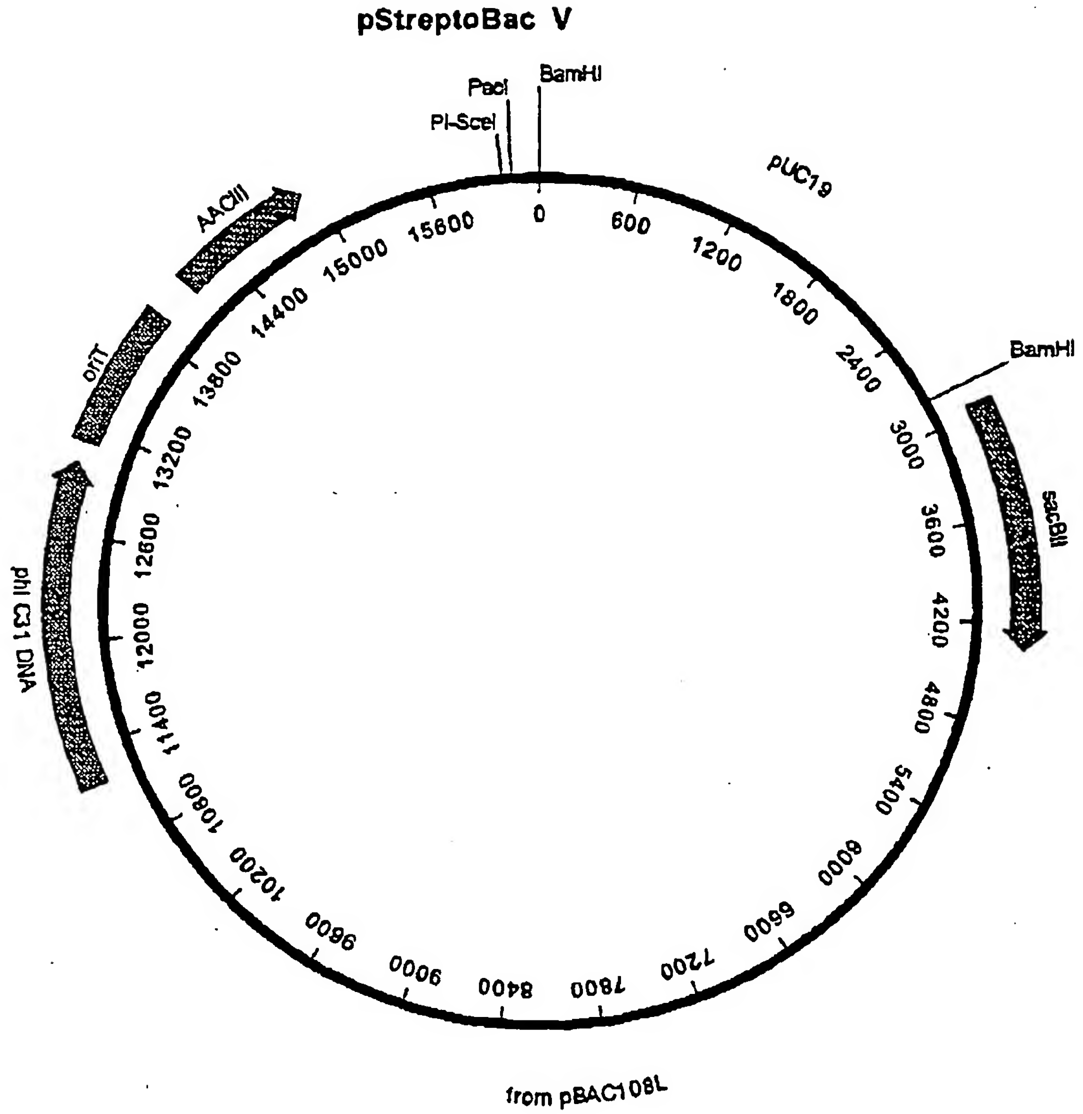
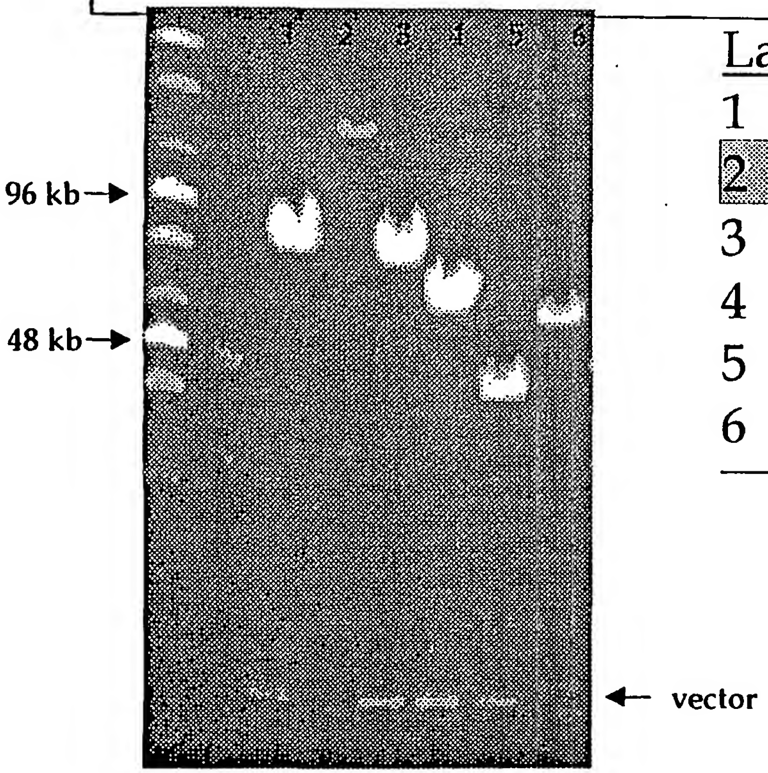

pStreptoBac V
BamHI
PacI
PI-SceI
pUC19
0
600
1200
1800
2400
3000
3600
4200
4800
5400
6000
6600
7200
7800
8400
9000
9600
10200
10800
11400
12000
12600
13200
13800
14400
15000
15600
BamHI
sacBII
from pBAC108L
phi C31 DNA
oriT
AACIII

Figure 7

Recovery of *dpt*-related clones



*Hin*DIII digest of BAC clones

Lane	BAC clone	Insert
1	01G05	82 kb
2	03A05	120 kb
3	06A12	85 kb
4	12F06	65 kb
5	18H04	46 kb
6	20C09	63 kb

Figure 8

BACs cover 180-200 kb in *dpt* region

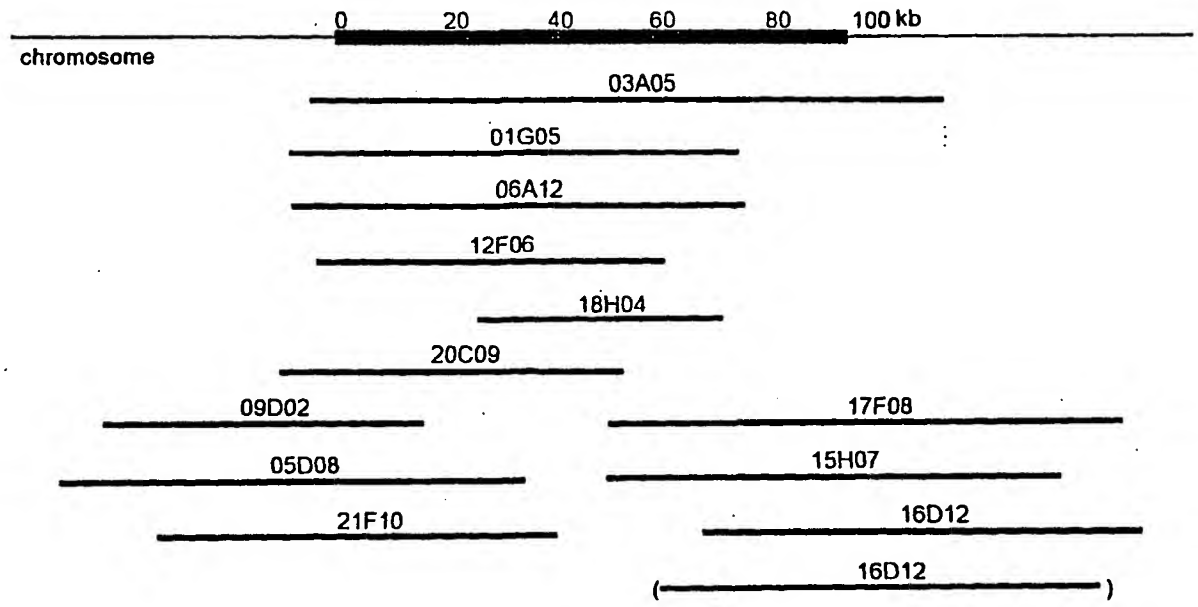
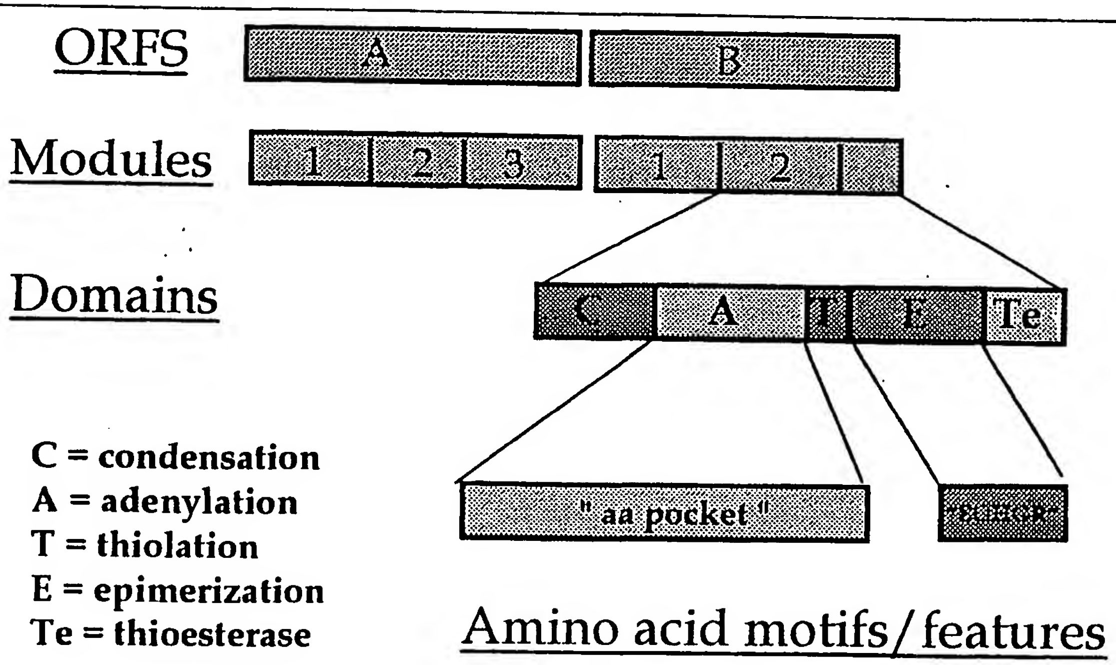


Figure 9

NRPS gene structure



C = condensation
A = adenylation
T = thiolation
E = epimerization
Te = thioesterase

Figure 10

A domain similarities (asn, asp)

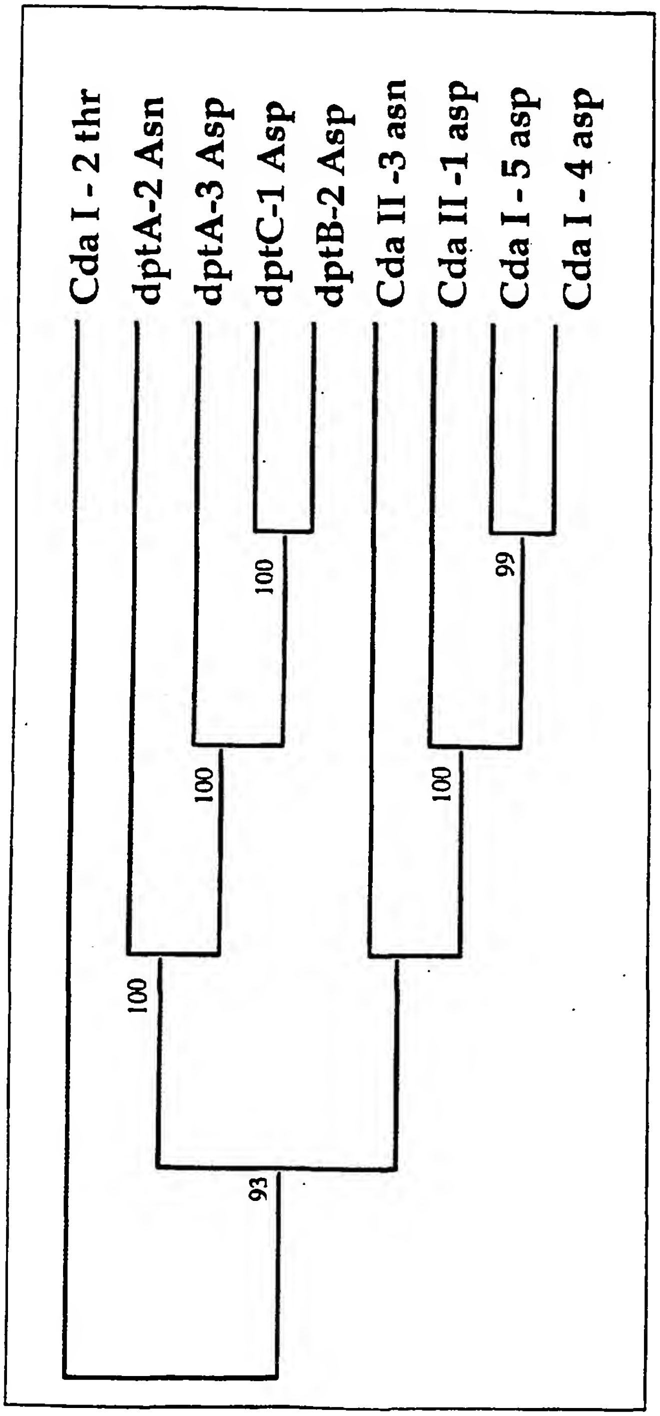
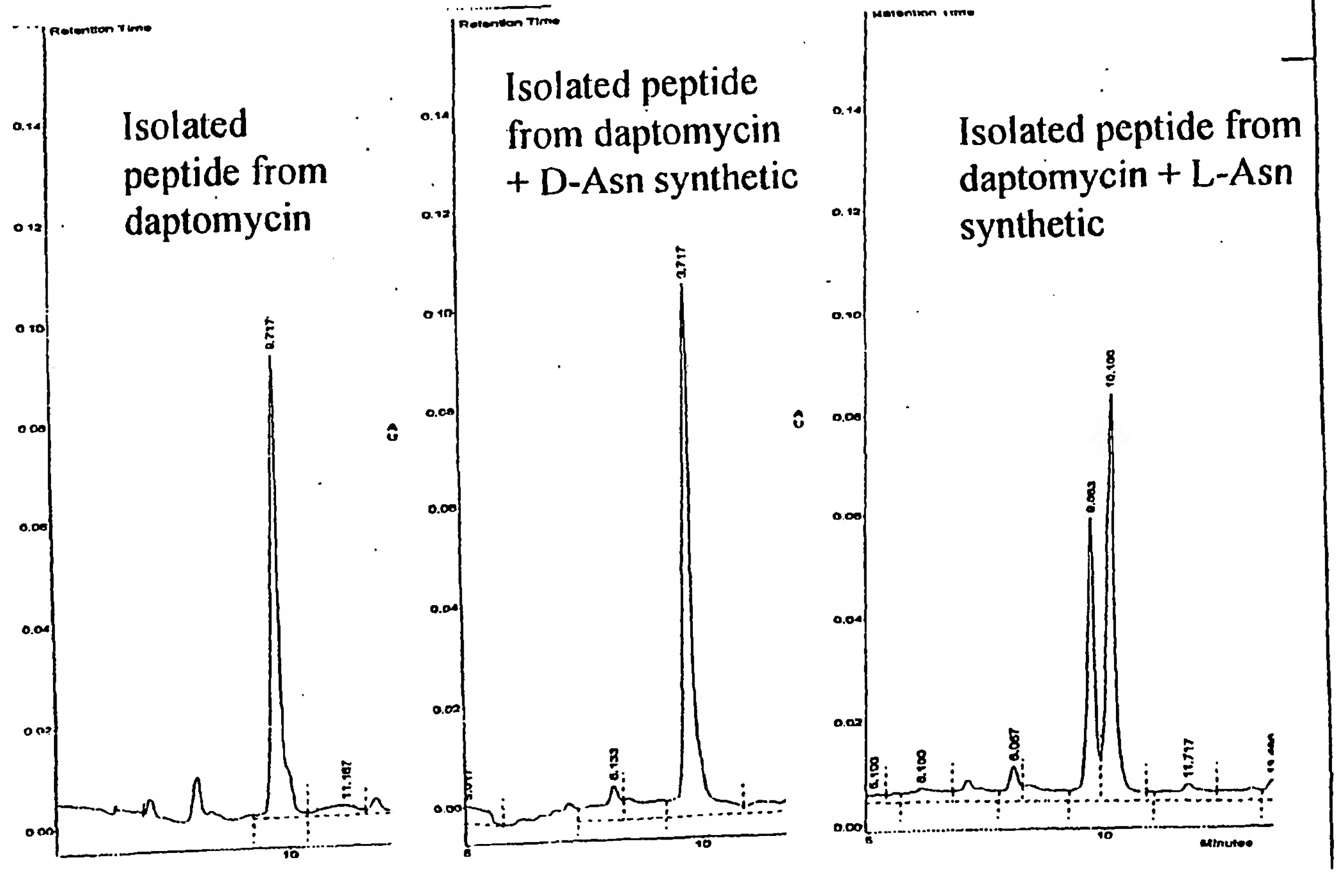


Figure 11

Stereochemistry of Asn

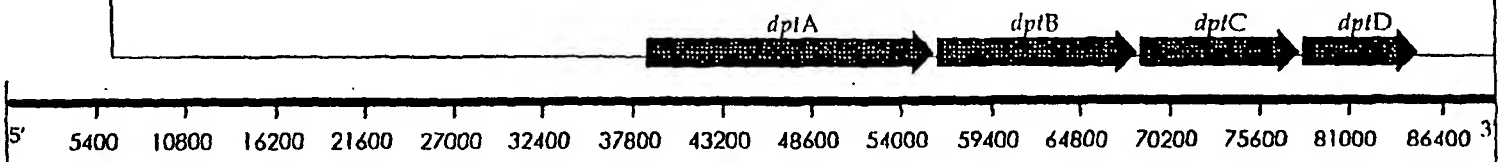

Isolated peptide from daptomycin

Isolated peptide from daptomycin + D-Asn synthetic

Isolated peptide from daptomycin + L-Asn synthetic

Figure 12

Organisation of 46.6 kb *dpt* region



ORF	modules	domains
dptA	5	CAT•CATE•CAT•CAT•CAT
dptB	3	CAT•CAT•CATE•C
dptC	3	AT•CAT•CATE
dptD	2	CAT•CATTe

4 ORFs translationally coupled
Epimerisation domains assocated with asn, ala and ser modules
One of the asp modules is split between *dptB* and *dptC*
Thioesterase (Te) domain at end of *dptD*